PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIAS DA COMPUTAÇÃO

EULANDA MARIA PEDRO DANIEL

UM MÉTODO DE REUTILIZAÇÃO DO PROCESSO DA DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS APLICADO NO SETOR AGRÍCOLA

Porto Alegre
2019

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM MÉTODO DE REUTILIZAÇÃO DO PROCESSO DA DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS APLICADO NO SETOR AGRÍCOLA

EULANDA MARIA PEDRO DANIEL

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Duncan Dubugras Alcoba Ruiz
Co-Orientador: Prof. Ana Paula Terra Bacelo

Porto Alegre
2019

Ficha Catalográfica

D184m Daniel, Eulanda Maria Pedro

Um Método de Reutilização do Processo da Descoberta de Conhecimento em Base de Dados aplicado no Setor Agrícola / Eulanda Maria Pedro Daniel . – 2019.

128 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.
Co-orientador: Prof. Dr. Ana Paula Terra Bacelo.

1. Processo da descoberta de conhecimento em BD agrícola. 2. Previsão do rendimento das culturas. 3. Mineração de dados. 4. Modelos preditivos. I. Ruiz, Duncan Dubugras Alcoba. II. Bacelo, Ana Paula Terra. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Salete Maria Sartori CRB-10/1363

Eulanda Maria Pedro Daniel

Um Método de Reutilização do Processo da Descoberta de Conhecimento em Base de Dados Aplicado no Setor Agrícola.

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciencia da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em 27 de setembro de 2019.

BANCA EXAMINADORA:

Profa. Dra. Renata de Matos Galante (PPGC/UFRGS)

Profa. Dra. Renata Vieira (PPGCC/PUCRS)

Profa. Dra. Ana Paula Terra Bacelo (PPGCC/PUCRS – Co-Orientadora)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho, em especial, à minha filha Kyanga, por tudo que passou durante a minha ausência física no período da minha formação. Dedico também, à minha mãe e minhas irmãs substitutas da minha filha, pelo apoio prestado durante a formação.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê”
(Arthur Schopenhauer)

AGRADECIMENTOS

À minha família e amigos, pelo apoio em todos os momentos da minha vida, em especial para minha filha Kyanga que, apesar da distância que nos separou, esteve ciente da causa e aguardou fortemente pelo meu regresso à casa, enfrentando todos os momentos difíceis pela a minha ausência.

À minhas irmãs, cunhados, em especial a minha mãe que nunca deixou de acreditar em mim. E olha, que ela rezou todos os dias, fazendo com que eu continuasse a acreditar em Deus para que chegasse até aqui. Agradeço a todos eles por terem cuidado, de forma incondicional, a minha filha durante a minha ausência.

Ao meu orientador, Prof. Dr. Duncan Dubugras Alcoba Ruiz, pelo ensino dispensado a mim durante a minha trajetória acadêmica. Ele sempre demonstrou muita capacidade técnica, sempre tentou extrair o melhor para este trabalho, me fazendo entender todos os conceitos que me faltavam. Pela sua paciência para me manter no foco desta dissertação e, muito particularmente pelas criticas construtivas para o desenvolvimento deste trabalho. Destaco um agradecimento muito especial pelo esforço que fez para conseguir prolongar o tempo de realização dessa pesquisa, para que terminasse com sucesso. Sem o qual, teria sido difícil materializar o meu sonho.

Aos colegas do laboratório GPIN, pelas críticas construtivas, pelas revisões ao trabalho, pelas sugestões dadas de forma incondicional e, sobretudo, por me terem ajudado a ultrapassar algumas dificuldades no ato do desenvolvimento deste trabalho.

À prof. Ana Paula Terra Bacelo que,juntamente com o prof. Dr Duncan, me orientaram para o desenvolvimento desta pesquisa. Caminhamos juntos, algumas vezes ultrapassando desafios de linguagem no mesmo idioma. Os contextos diferentes fazem com que alguns termos semelhantes utilizados no Brasil tenham um significado diferente do Português falado em Moçambique.

Um grande agradecimento ao Ministério da Ciência e Tecnologia, Ensino Superior e Técnico Profissional (MCTESTP) por financiar a minha formação, me oferecendo uma oportunidade para realizar o meu sonho de me tornar mestre em ciência da computação.

À todos que torceram por mim, muito obrigado.

UM MÉTODO DE REUTILIZAÇÃO DO PROCESSO DA DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS APLICADO NO SETOR AGRÍCOLA

RESUMO

A presente dissertação é desenvolvida com base na metodologia proposta por Fayyad, Piatetsky-Shapiro e Smyth, que se resume na extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um conjunto de dados. Partindo desse pressuposto, desenvolveu-se a presente dissertação em que se apresentam provas de conceito, tendo como objetivo agilizar o processo de descoberta de conhecimento em base de dados no setor agrícola, a partir do reuso das tarefas do processo. Este processo usa um conjunto de dados que, na maioria das vezes, necessita de ajustes ou configurações durante a execução do processo de Descoberta de Conhecimento em Base de Dados (DCBD), o que torna um problema devido a repetição das configurações em processos similares. Esta tarefa, torna a atividade exaustiva para o especialista de domínio ao ter que repetir toda a configuração em um novo processo, gastando um tempo que se poderia aproveitar utilizando uma base de conhecimento do aprendizado da execução prévia - um guia para a reutilização do processo de descoberta de conhecimento em base de dados contendo configurações necessárias das execuções de um processo prévio inicial, utilizado sobre a cultura de arroz. Em conclusão, no aproveitamento dos testes executados, nessa cultura, manifestaram-se em resultados positivos na execução da cultura do feijão, provando a premissa da reutilização do processo de descoberta de conhecimento em base de dados em outras culturas através das análises qualitativas e quantitativas feitas sobre os processos. Nos experimentos executados, o método constituído pela base do conhecimento do aprendizado, teve um ganho relativo da duração da execução do processo em 42,37% mais ágil em comparação com a execução do processo DCBD executado. A ferramenta WEKA foi a usada para a execução do processo, recorrendo, como nos referimos antes, a metodologia proposta por Fayyad, Piatetsky-Shapiro e Smyth, como referência teórica extremamente fundamental.

Palavras-Chave: Processo da descoberta de conhecimento em base de dados agrícola, previsão do rendimento das culturas, mineração de dados, modelos preditivos.

A METHOD OF REUSING THE KNOWLEDGE DISCOVERY PROCESS APPLIED TO THE AGRICULTURAL SECTOR

ABSTRACT

The current dissertation was developed based on the methodology proposed by Fayyad, Piatetsky-saphiro, and Smyth, which consists of the extraction of implicit information, previously unknown and potentially useful, from stored data. From that assumption, we developed the current work, where we present proof of concept, with the primary objective of speeding up the process of knowledge discovery using data from the agricultural sector from the reuse of task from the process. This process uses a set of data that most of the time requires adjustments or configurations during the execution of the knowledge discovery in databases (KDD) method, which creates a problem given the repetition of the settings in similar processes. This task turns the activity exhaustive for the domain specialist, by having to repeat the entire configuration in a new method, spending the time that could be leveraged using a pre-execution learning knowledge — a guide to reuse of the knowledge discovery process in a database containing necessary settings for executing an initial pre-process, used on rice cultivation. In conclusion, in taking advantage of the tests performed in this culture have shown positive results in the bean crop, proving the premise of the process reuse knowledge discovery in other cultures. In the performed experiments, the method based on the knowledge of learning, had a relative gain of the duration of the process execution by 42.37 % more agile compared to the execution of the executed DCBD process. The tool WEKA was the one used for the execution of the process, using, as in the referred earlier to the methodology proposed by Fayyad, Piatetsky-Shapiro, and Smyth as an extremely fundamental theoretical reference.

Keywords: Knowledge Discovery Process in Agricultural Database, crop yield prediction, data mining, predictive models.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AR – Additive Regression

BG – Bagging

DCBD – Descoberta de Conhecimento em Base de Dados

DLA – Distância Latitudinal

KDD – Knowledge discovery in databases

KNN – K-Nearest Neighbors

LR – Linear Regression

MAE – Erro Absoluto Médio

MLP – Multilayer Perceptron

M5P – M5-Prime

R – Coeficiente da Correlação

REPTREE – Reduced-Error Pruning Tree

RBD – Regression By Discretization

RMSE – Raiz do Erro Médio Quadrático

RT – Random Forest

VC – Validação Cruzada

WEKA – Waikato Environment for Knowledge Analysis

SUMÁRIO

1. INTRODUÇÃO

Hoje em dia muitas empresas agrícolas necessitam entender o conhecimento extraído no conjunto de dados produzidos para melhor interpretação do conhecimento oculto que eles armazenam. O conhecimento extraído antecipa tendências e comportamentos futuros para permitir que as empresas agrícolas tomem decisões sensatas orientadas pelo conhecimento obtido.

O processo de Descoberta de Conhecimento em Bases de Dados (DCBD) é uma área de pesquisa estabelecida que ocupa cada vez mais o papel central no setor agrícola e, é responsável por extrair o conhecimento oculto existente nos conjuntos dos dados. Segundo Veenadhari *et al.* [VMS11], os conjuntos de dados gerados no setor agrícola facilitam as numerosas tarefas de predição do rendimento através de informações mais cruciais encontradas durante a execução do processo.

O problema do processo de DCBD é comum em todas as áreas de aplicação do processo. Este processo não possui uma parametrização que mantém as configurações executadas de um processo anterior, e também não possui uma base de conhecimento do aprendido para usar como guia de implementação em novas execuções do processo de DCBD, podendo ser capaz de prever o rendimento das culturas.

Este tipo de pesquisa ajuda o setor agrícola no planejamento de suas atividades procurando evitar, assim, surpresas desagradáveis durante a plantação ou colheitas das culturas. Segundo Machado *et al.* [MPdL+17], o setor responsável pode usar o processo para tomar decisão certa a partir dos resultados obtidos com a extração do conhecimento gerado.

Durante a sua execução, o processo opera apenas para um único conjunto de dados. Segundo Priyadharsini [PT14], a atividade torna o trabalho do analista custosa devido a repetição de tarefas configuradas, o que torna um problema quando se deseja executar o processo em diferentes conjuntos de dados num mesmo domínio.

A parametrização pode ser disponibilizada para o uso, através de uma base de conhecimento gerada que disponibiliza as configurações aprendidas do processo prévio, que o mesmo servirá como guia para reutilização em novos processos similares. Essa base tem como finalidade agilizar a execução do processo através da reutilização das configurações catalogadas na base de conhecimento.

Devido a este problema, surge a motivação do estudo cujo objetivo centra-se em desenvolver um método baseado em construção de uma base de conhecimento com parametrizações aprendidas a partir da execução de um processo base inicial. A partir desta base, um usuário poderá reutilizar em novos processos de DCBD em um domínio de culturas temporárias. A proposta do estudo poderá tornar ágil a execução do processo de

DCBD no setor agrícola por meio do reúso da base de conhecimento gerada, testada sobre análises qualitativas e quantitativa em relação a duração do processamento do método.

1.1 CONTEXTUALIZAÇÃO

Nos dias de hoje, o processo de DCBD exige que o analista defina todas as transformações necessárias nos dados para que consiga extrair o máximo de conhecimento possível. Essas configurações são repetidas quando se aplica a um outro processo do mesmo domínio. O repetir significa obrigar que este processo invista tempo no mapeamento dos novos atributos e na aplicação dos ajustes ou configurações que foram considerados necessários no processo anteriormente definido.

A realização de operações no processo sobre mais do que um conjunto de dados, segundo Zhang *et al.* [ZWZ03], acarreta problemas na sua execução. Os autores acrescentam ainda que a realização das operações sobre múltiplos conjuntos de dados devem ser feitas a nível local para evitar diferenciação nas configurações. Ou por outra, devem ser criados processos de DCBD isoladamente para cada conjunto de dados e não serem feitas análises conjuntas à globalidade dos dados.

Essas análises conjuntas acabam por resultar em desgaste para os especialistas e custos para a realização do processo devido a morosidade no processamento e pelas repetições encontradas durante a realização das tarefas do processo. Acontece que, segundo Ruping *et al.* [RWB10], existem muitas análises que são feitas sobre os dados mas que não podem ser compartilhadas, por vezes, entre configurações semelhantes mas não rigorosamente iguais pertencentes a um mesmo domínio.

Essas situações acontecem porque os conjuntos de dados selecionados para o processo nem sempre possuem os mesmos atributos, ou um mesmo atributo encontra-se representado de modo diferente, ou até mesmo porque possui atributos que não são utilizados ou que necessitam de ajuste. Por vezes, o tratamento dos dados faz com que o especialista realize determinada análise sobre um conjunto de dados e que tenha que recomeçar todo o processo de DCBD quando executado em outro conjunto de dados do mesmo domínio. Por exemplo, o especialista constrói um processo de DCBD em que ajusta e configura diferentes tarefas sobre o processo.

Entretanto, durante a execução de um novo processo, um especialista normalmente executa todo o processo novamente como se fosse a sua primeira implementação e com isso ele despende mais tempo na reconfiguração do novo processo que seria útil para aplicar a mesma configuração sobre o novo conjunto de dados, como os autores Wegener *et al.* [WR10] argumentam.

1.2 MOTIVAÇÃO

O avanço no uso do processo de DCBD na agricultura tem se expandindo devido a elevada quantidade de dados gerados pelo setor, quando se pretende prever o rendimento das culturas. Visando melhorar a qualidade da produção através de um bom planejamento das atividades agrícolas, este processo é executado por especialistas do domínio que, de acordo com Garcia e Camolesi [GCJ15], este processo é realizado de forma unária sobre as culturas.

Segundo Shepher *et al.* [SP15], realizar um estudo que possibilite melhoria e aumento no rendimento de culturas por meios tecnológicos pode reduzir os custos para o setor agrícola. Devido a importância que a agricultura possui, a nível mundial, e no rendimento que a produção agrícola traz para cada país, é vantajosa a execução de um processo de DCBD.

El-Sappagh *et al.* [ESEMRE13], argumentam que o processo de DCBD é um processo executado sobre diferentes bases de dados em diferentes áreas de aplicação, permitindo extrair padrões presentes nos mesmos, mas que essa execução é realizada em uma base de dados para cada execução do processo. Essas ações tornam as atividades custosas, pois são executadas repetitivamente em cada conjunto de dados.

Quando o processo é executado sobre um conjunto de dados de uma cultura e em seguida é necessário executar um outro processo sobre um novo conjunto de dados no mesmo domínio é verificado que não existe um catálogo com configurações exercidas sobre o processo. Sem esse catálogo, os novos processos a serem executados sobre esse domínio tendem a ser executados como se fosse a primeira vez, realizando repetitivamente todas as tarefas executadas no processo anterior.

Este problema da repetibilidade da execução, adaptação e ajustes de tarefas, motivou na possibilidade da construção de uma metodologia aplicada no setor agrícola que possa agilizar o processo de DCBD com a finalidade de prever o rendimento de diferentes culturas temporárias, através da reutilização do processo de DCBD agrícolas.

A metodologia consiste na criação de base de conhecimento contendo todo o aprendizado do processo, permitindo a sua reutilização e evitando, assim, o retrabalho de configuração. A base de conhecimento tem função de servir como guia para a execução de um processo de DCBD agrícola, onde todas as parametrizações suportadas pelo processo estão catalogadas para um reúso em novo processo similar.

1.3 OBJETIVOS

Neste trabalho propõe-se uma metodologia para reutilização do processo de descoberta de conhecimento em base de dados agrícolas implementadas em um domínio de culturas temporárias com intuito de prever o seu rendimento.

Assim sendo, o trabalho objetiva essencialmente:

- Desenvolver uma base de conhecimento através do apreendido de um processo de DCBD para reutilizar em novos processos;
- Definir configurações realizadas do processo de DCBD prévio;
- Testar o método desenvolvido mediante a execução de um novo processo com apoio da base de conhecimento;
- Testar a solução desenvolvida mediante as análises qualitativas e quantitativas dos processos executados.

Através destes objetivos espera-se minimizar o retrabalho das tarefas e agilizar a execução do processo de DCBD aplicado no setor agrícola.

1.4 JUSTIFICATIVA

As contribuições deste trabalho são as seguintes:

- Servir de fonte para realizar novos processos de DCBD na agricultura sobre novos conjuntos de dados, visto que o assunto é relativamente novo e não existe literatura que se assemelhe ao assunto em estudo na presente dissertação;
- O desenvolvimento do método da pesquisa envolve práticas que podem ser aproveitadas em áreas diversas, pois oferece perspectivas para o desenvolvimento de outros projetos automatizados.
- A utilização de ferramenta como WEKA e algoritmos de previsão numérica na mineração de dados com resultados práticos sobre um estudo de caso, pode servir como fator motivacional para a implementação de novos modelos nas pesquisas;
- Para o setor agrícola, este trabalho poderá ser utilizado em ações práticas, favorecendo o processo de tomada de decisão, direcionando estratégias e processos. Além disso, o conhecimento oculto a ser descoberto nos padrões minerados pode ser utilizado como vantagem competitiva e aperfeiçoamento das ações de relacionamento com outras empresas agrícolas;

- Em termos de inovação, esta dissertação apresenta um método de reutilização do processo de DCBD agrícola executado na cultura do arroz como processo prévio e testado na cultura do feijão.

A literatura encontrada referenciando o processo de DCBD na agricultura trata da aplicação para a previsão da produção e rendimento das culturas. Assim, a configuração do modelo para a reutilização do processo de DCBD para a previsão do rendimento das culturas é distinta e diferenciada das publicações e referências disponíveis.

1.5 METODOLOGIA

Este trabalho caracteriza-se, em termos metodológicos, como sendo uma execução descritiva sobre a forma de conhecimento aprendido de um processo prévio executado. Propõe o método que permite a reutilização do processo de DCBD no setor agrícola, por meio de geração de uma base de conhecimento aprendido a partir das tarefas realizadas, favorecendo uma posterior tomada de decisão e formação de ações ágeis durante a execução de novos processos.

Metodologicamente, o trabalho trata de um processo de regressão para a previsão do rendimento das culturas, avaliando as principais técnicas de DCBD aplicadas em culturas temporárias. As atividades envolvidas na elaboração deste trabalho são descritas em função do método de descoberta de conhecimento em base de dados proposta por Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth em seus estudos.

O método proposto do processo de DCBD ou KDD, sigla em inglês, proveniente de *Discovery Knowledge in Databases* é um processo, segundo FAYYAD *et al.* [FPSS96], de várias etapas não triviais, interativo e iterativo, que extrai padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados, em que a extração de conhecimento é um processo dinâmico e evolutivo. Os detalhes da execução deste método serão encontrados nos capítulos 3.

1.6 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está dividido em 5 capítulos, organizados da seguinte maneira:

- Capitulo 1 apresenta a introdução, motivação e objetivo do trabalho.
- Capítulo 2 apresenta a fundamentação teórica da pesquisa.
- Capítulo 3 apresenta o modelo aplicado para a realização do estudo empírico.

- Capítulo 4 apresenta o teste da solução do método.
- Capítulo 5 apresenta as considerações finais do trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos que serão aplicados no decorrer deste trabalho, enunciando a importância e o conceito sobre o processo de DCBD num contexto geral e agrícola. São apresentados de forma sucinta, nas suas seções, as técnicas, os algoritmos de predição e as principais técnicas de avaliação de desempenho usadas na agricultura, que fazem ligação ao objetivo principal do estudo, a reutilização do processo de DCBD no setor agrícola para a previsão do rendimento das culturas. Definiremos também, a ferramenta base usada durante a execução do método.

2.1 REUTILIZAÇÃO EM DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Nos dias de hoje existem muitas análises que são feitas sobre os dados agrícolas mas que não podem ser partilhadas, por vezes, entre configurações semelhantes mas não rigorosamente iguais pertencentes a um mesmo domínio. Essas situações acontecem porque os conjuntos de dados selecionados para o processo de DCBD nem sempre possuem os mesmos atributos ou o atributo encontra-se representado do modo diferente, ou até mesmo porque possui atributos que não são utilizados ou que necessitam de serem ajustados. Para além do mesmo, esses dados sofrem algumas configurações durante a execução do processo de DCBD que também não são partilhadas o que faz com que haja retrabalho durante a execução de um novo processo

A ideia do reuso é evitar retrabalho no desenvolvimento de um novo processo, sempre levando em consideração trabalhos anteriores, fazendo com que soluções previamente desenvolvidas sejam aproveitadas e executadas em novos contextos. Uma possível solução, que pode não resolver todos os problemas, mas pode ajudar a lidar com esse processo é a reutilização.

A reutilização em engenharia de software, segundo Pacheco *et al.* [PGCMA15], se baseia no uso de conceitos, produtos ou soluções previamente elaboradas ou adquiridas para criação de um novo software, visando melhorar significativamente a qualidade e a produtividade. Segundo Basha e Mohan [BM14], Reusar um produto significa poder reusar partes de um sistema desenvolvido anteriormente como: especificações, módulos de um projeto, arquitetura e código fonte.

A principal motivação para a reutilização em engenharia de software está relacionada ao aumento dos níveis de qualidade e produtividade no desenvolvimento de software. Neste sentido, este trabalho apresentará o conceito de reutilização como um método que utiliza um processo de DCBD para reutilização das tarefas executadas para aplicar em novos processos. A vantagem de reutilizar esse processo é gerenciar a complexidade da

execução do processo, evitando o retrabalho de configuração, aplicando algoritmos conhecidos para o problema de regressão e agilizar o processo de DCBD no setor agrícola.

No caso do presente trabalho, para reutilizar o processo de DCBD é necessário refazer as fases do DCBD que são necessárias e aproveitar assim as restantes fases do processo anterior utilizando uma base de conhecimento que contém as configurações essências a serem reutilizadas. Este método permite otimizar e agilizar as tarefas desenvolvidas pelos analistas, para que estes não despendam demasiado tempo na configuração e na transformação dos dados, rentabilizando esse mesmo tempo na extração de conhecimento ou em outras atividades que possam ser necessárias no processo.

2.2 PROCESSO DA DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Na atualidade, as empresas geram grandes quantidades de dados devido a grandes tecnologias de ponta existentes. Esse crescimento na quantidade de dados armazenados, gerou a necessidade por novas técnicas e ferramentas automatizadas que podem auxiliar na transformação desses dados em informação útil e conhecimento que pode ser utilizados para vários fins.

As ferramentas de análise são necessárias quando há abundância dos dados. A situação é conhecida como “rica em dados, mas pobre em informação”. Segundo Han e Kamber [HPK11], existe pouca informação nos conjuntos de dados devido a falta de execução de métodos que possam ajudar a extrair a informação ou ainda o conhecimento que eles carregam muitas vezes não interpretados pelo homem.

Existem várias metodologias de DCBD propostas por diversos autores, mas neste trabalho apresentaremos a metodologia proposta por Fayyad, Piatetsky-Shapiro e Smyth por ser a mais conhecida na área de DCBD, e que foi definida na seção 1.5. Outros autores como Goldschmidt *et al.* [GP05], argumentam que o DCBD é importante para a tomada de decisão pois permite a aplicação de técnicas para auxiliar em qualquer processo de descoberta de conhecimento em base de dados.

Os autores como Dos Santos *et al.*[dSSdL+16] argumentam que existem três propriedades que o conhecimento a ser descoberto deve satisfazer: ser o mais correto possível; ser compreensível; e ser interessante, útil e/ou novo. Este processo de DCBD tem como finalidade a extração de conhecimento útil dos seus detentores.

Fayyad *et al.* [FPSS96] explicam que o processo é interativo porque há interferência humana na interpretação e na tomada de decisão, e iterativo porque pode haver repetições em todo processo ou em alguma das etapas que o compõe. Quanto ao não trivial, o autor Boente [BGE08] faz alusão da complexidade na execução que o processo

pode oferecer. Na Figura 2.1 é apresentada a sequência das fases do processo DCBD a ser seguida no desenvolvimento do estudo.

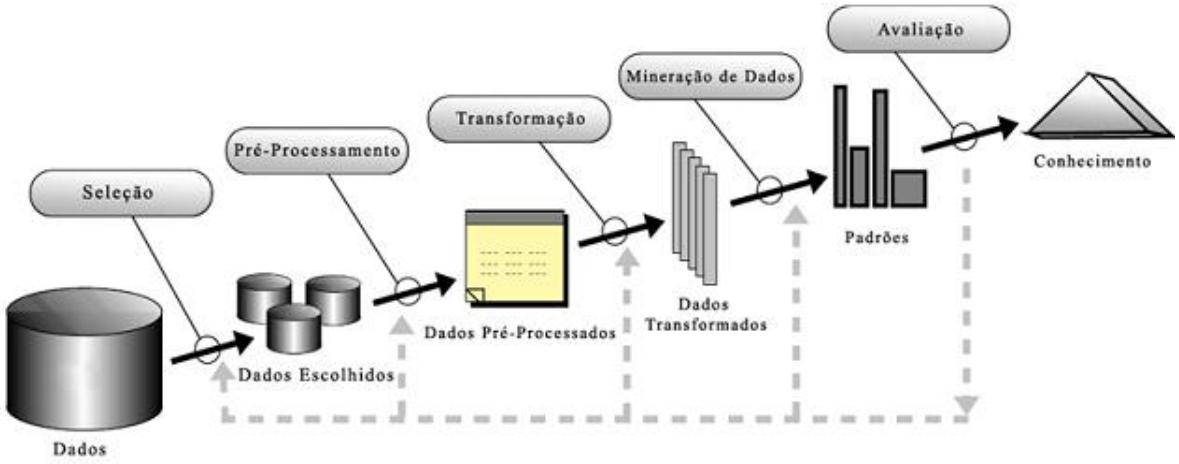


Figura 2.1: Processo de DCBD proposta por Fayyad *et al.* [FPSS96]

As etapas do processo ilustrado na figura 2.1 levam-nos ao alcance do objetivo. A fase inicial, a **Seleção**, diz respeito a quais dados serão utilizados nas próximas etapas. Nesta etapa é necessário que o executor do processo possua os objetivos definidos e o tipo de conhecimento que deseja extrair da base de dados.

O resultado final da aplicação do processo DCBD depende inicialmente desta etapa, pois a saída obtida vai depender de quais dados foram selecionados porém nem todos os dados contidos nas bases de dados influenciam positivamente no objetivo do processo. Esta etapa consiste em selecionar um conjunto ou subconjunto de dados que farão parte da análise. As fontes de dados podem ser variadas tais como planilhas, sistemas gerenciais, data warehouses e podem possuir dados com formatos diferentes.

A fase do **Pré-processamento** consiste em verificar a qualidade dos dados armazenados. A base passa por um processo de limpeza, correção ou remoção de dados inconsistentes, além de verificar dados ausentes ou incompletos e identificar anomalias.

A fase da **Transformação** consiste em aplicar técnicas de transformação tais como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Aqui os dados ficam disponíveis e agrupados em um mesmo local para a aplicação dos modelos de análise.

A fase da **Mineração de Dados** consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo verificar uma hipótese, e descobrir novos padrões de forma autônoma. Além disso, a descoberta pode ser dividida em: preditiva e descritiva, definidos na seção 2.3. Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.

A fase da **Interpretação e Avaliação** é a ultima etapa do processo que consiste em avaliar o desempenho do modelo aplicado em cima dos dados que não foram utilizados na fase de treinamento ou mineração. A validação pode ser feita de diversas formas, algumas delas utilizam medidas estatísticas, como o caso em estudo.

2.3 PROCESSO DA DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS NA AGRICULTURA

Nos estudos sobre o processo aplicado na agricultura é possível encontrar trabalhos que relacionam o processo de DCBD com a mineração de dados. O processo de DCBD refere-se ao processo geral de descoberta de conhecimento em base de dados, enquanto que a mineração de dados se refere a uma fase específica do processo de DCBD que extrai padrões do conjunto dos dados.

Segundo Kamilaris *et al.* [KKPB17], a mineração de dados oferece grande vantagem na agricultura para detecção de doenças, previsão de problemas e otimização de pesticidas. Acrescentam ainda os autores Kodeeshwari e Ilakkiya [KI17], que a mineração de dados também é usada para mostrar as informações estatísticas para prever a safra futura, previsão do tempo, pesticidas e fertilizantes a serem usados, receita a ser gerada e assim por diante.

Surya e Aroquiaraj [SA18] acrescentam que na agricultura, a previsão envolve o uso de algumas variáveis ou campos no banco de dados para prever valores desconhecidos e futuros de outras variáveis de interesse. Através deste processo, as organizações são capazes de produzir informações preditivas como suporte para a tomada de decisão se a conclusão adquirida for aceitável para os decisores.

Mucherino e Ruß [MR11], afirmam que a previsão sobre o rendimento da produção pode ser, por exemplo, resolvida com técnicas de mineração de dados preditivos. Para isso, deve-se considerar, no processo, que existem dados de algum tempo passado para os quais o rendimento de produção apropriado foi anotado.

Todas essas informações criam um conjunto de dados que podem ser usados para aprender formas de predizer os rendimentos futuros da produção, porque os novos dados do conjunto estão disponíveis. Neste contexto, o setor agrícola precisa utilizar técnicas de mineração preditiva para predizer eventos futuros.

2.3.1 TÉCNICAS DE MINERAÇÃO PREDITIVAS NA AGRICULTURA

Devido os avanços da tecnologia de geração e armazenamento de dados, os setores passaram a acumular grandes quantidades de dados. Extrair informação útil deste contexto pode ser extremamente desafiador. Witten *et al.* [WFHP16], definem a mineração de dados como sendo a extração de conhecimento implícito, previamente desconhecido e potencialmente útil a partir dos dados e com o auxílio de modelos preditivos.

Segundo Moor [MHN15] um modelo preditivo possui como objetivo prever o valor de um atributo em particular, baseado no valor de outros atributos. O atributo a ser previsto é normalmente conhecido como **atributo alvo** ou **variável dependente**, enquanto os atributos utilizados para fazer a previsão são conhecidos como **explanatórios** ou **variáveis independentes**.

Nesta dissertação, a modelagem preditiva se refere à tarefa de construir uma função que modele o atributo alvo construindo um estimador. O objetivo do aprendizado preditivo é aprender uma função que mapeia as variáveis independentes no atributo alvo. Se o conjunto de dados for composto de valores nominais, tem-se um problema de classificação, ou aprendizado de conceitos e o estimador é um classificador. Se o domínio for um conjunto infinito e ordenado de valores, tem-se um problema de regressão, que induz um regressor, Faceli *et al.* [FLG+11].

Nas Seções 2.3.1.1 e 2.3.1.2 são apresentados exemplos de modelos preditivos de classificação e regressão usados na agricultura e alguns utilizados em nosso trabalho. No nosso estudo, estes modelos são usados para prever o rendimento das cultura, onde são comparados os resultados obtidos das métricas dos erros com objetivo de encontrar o modelo com melhor desempenho para o problema. Para além disso, é sugerido neste trabalho os modelos que podem ser utilizados para resolver o problema de previsão de culturas baseando em maior frequência entre os modelos selecionados que tiveram um bom desempenho ao longo dos testes executados.

2.3.1.1 PREDIÇÃO BASEADA EM CLASSIFICAÇÃO

De acordo com Kesavaraj *et al.* [KS13] classificação é um processo de encontrar a função para classificar os dados em uma das várias classes. Para tarefas de classificação, a variável alvo (objetivo) geralmente tem um pequeno número de valores discretos. Ela pode ser usada para extrair modelos que descrevem classes de dados importantes ou para prever tendências futuras de dados.

Faceli *et al.* [FLG+11] afirma que na classificação, o modelo aprende a prever um rótulo de classe a partir de um conjunto de dados de treinamento que pode ser usado para prever rótulos de classes discretas em novas amostras. Um dos principais objetivos do

classificador é maximizar a precisão preditiva obtida pelo modelo ao classificar exemplos no conjunto de testes não vistos durante o treinamento.

2.3.1.2 PREDIÇÃO BASEADA EM REGRESSÃO

Os autores Fetanat *et al.* [FMZ15] definem a Regressão como sendo uma técnica de mineração de dados (aprendizado de máquina) usada para ajustar uma equação a um conjunto de dados. A forma mais simples de regressão, de acordo com os autores, é a regressão linear que usa a fórmula de uma linha reta ($y = mx + b$) e determina os valores apropriados para m e b para predizer o valor de y com base em um dado valor de x.

O objetivo da regressão é encontrar uma função de mapeamento que pode ajustar os dados de entrada minimizando o erro. Suponha-se que um analista no setor agrícola esteja interessado em prever o rendimento das culturas. Esta análise de dados é um exemplo de predição numérica em que o modelo construído irá prever uma função de valor contínuo. Este modelo é um preditor. Segundo Han *et al.* [HPK11], a análise de regressão é uma metodologia estatística que é amplamente utilizada para previsão de números. Por se tratar de um preditor, esta metodologia é aplicada no neste cenário em estudo.

2.3.2 ALGORITMOS PREDITIVOS NA AGRICULTURA

Na agricultura são frequentes pesquisas de natureza preditiva onde diversos autores aplicam diferentes modelos com objetivo de prever um dado valor continuo ou uma classe. Embora citamos o uso dos algoritmos na agricultura, estes também podem ser aplicados em diversas áreas desde que o objetivo seja prever um dado valor.

Os algoritmos mais comuns aplicados na agricultura são os apresentados na tabela 2.1, que podem ser usados tanto como classificadores assim como regressores. Apresentamos também a finalidade da aplicação no setor agrícola.

Tabela 2.1: Algoritmos de predição usados na agricultura

Autores	Algoritmos preditivos	Aplicação
Dey *et al.* [DMU17]	Multiple Linear Regression, AdaBoost (Adaptive Boosting), Support Vector Machine, Regression Linear e Modified Nonlinear Regression (MNR).	Desenvolvem um sistema que prevê o rendimento do arroz usando as condições climáticas da região de Bangladesh na Índia.
Diriba e Borena [DB13]	J48, Random Forest e REPTree.	Avaliam aplicações preditivas de mineração de dados que podem ser aplicados sobre a produtividade das culturas agrícolas da Etiópia, concentrando-se em pequenos agricultores.
Gonzalez *et al.* [SAS+14a]	Multiple Linear Regression, Stepwise Linear Regression, M5 Regression Trees e Multilayer Preceptron.	Avaliam as técnicas mais comuns de modelagem orientada a dados aplicados para prever o desempenho do rendimento das culturas, usando um método completo para definir o melhor subconjunto de atributos para cada modelo.
Ruß e Kruse [RK10]	Support Vector Regression, Random Forests e Bagging.	Melhoram os modelos existentes de previsão de rendimentos e, além disso, incorporam uma metodologia genérica de agrupamento espacial no processo.
Gonzalez *et al.* [SAS+14b]	Multiple Linear Regression, M5-Prime Regression Trees, Perceptron Multilayer Neural Networks, Support Vector Regression e K-Nearest Neighbor.	Comparam a precisão preditiva de aprendizado de máquina e técnicas de regressão linear para predição de rendimento das culturas.
Pudumalar *et al.* [PRR+17]	Random Tree, CHAID, K-Nearest Neighbor e Naive Bayes.	Propõem um sistema de recomendação. A partir de parâmetros específicos do local, utilizam conjuntos de modelos com técnicas de votação por maioria para recomendar uma cultura.
Geetha e Elizabeth [GS18]	Regression By Discretization	Prevê o perfil do solo por meio de regressão modificada pelo algoritmo de discretização (RBD) para o rendimento da cultura no distrito de Trichy na Índia.
Bhojani e Bhatt [BB18]	Gaussian Processes, Multilayer Perceptron, Kstar, Sequential Minimal Optimization, M5Rules e Additive Regression.	Aplica técnicas de mineração de dados para prever o rendimento da safra de trigo nos distritos do Estado de Gujarat na Índia.
Sellam e Poovammal [SP16]	Linear Regression.	Analisam parâmetros ambientais que influenciam o rendimento da lavoura e estabelecem uma relação entre esses parâmetros.

De acordo com Faceli *et al.* [FLG+11], na aplicação dos algoritmos a problemas reais, o conhecimento que se tem do domínio a ser investigado é provido unicamente pelo conjunto de exemplos a partir do qual a indução de um modelo preditivo é então realizada. Não se pode afirmar que existe uma técnica que se sairá melhor no desempenho da resolução de qualquer tipo de problema.

Há necessidade de experimentação controlada para validar qualquer técnica proposta em que se demonstre a sua efetividade na solução de diferentes problemas. Dentre os algoritmos apresentados pelos autores, com diferentes aplicações, são definidos dez algoritmos de predição baseados em regressão apresentados com objetivo de prever o rendimento das culturas e que os mesmos serão aplicados no presente estudo.

1. *Reduced-Error Pruning Tree* (REPTree): É um algoritmo de aprendizado de árvore de decisão rápida, usado tanto para a classificação quanto para a regressão. Baseia-se no princípio de calcular o ganho de informação com a entropia e minimiza o erro resultante da variação. Diriba e Borena [DB13] aplicam RepTree de regressão para gerar várias árvores em iterações alteradas. No conjunto de dados do setor agrícola, o método torna-se viável quando usado para ganhar informação com a melhor de todas as árvores geradas pelo algoritmo. Ele classifica os valores dos atributos numéricos.

2. *Random Tree* (RT): É uma árvore aleatória em que para cada divisão apenas um subconjunto aleatório de atributos está disponível. Pudumalar *et al.* [PRR+17] usaram o algoritmo para dados nominais prevendo o rótulo da classe para cada um dos conjuntos de dados de treinamento. Mas os autores afirmam que da mesma forma, podem ser usados para dados numéricos.Sendo assim, este algoritmo será aplicado em nosso conjunto de dados para predizer o rendimento das culturas, considerando que a variável alvo é um dado numérico.

3. *Random Forest* (RF): É uma combinação de preditores de árvores, de tal forma que cada árvore depende dos valores de um vetor aleatório amostrado de forma dependente e com a mesma distribuição para todas as árvores na floresta. Na versão utilizada pelos autores Ruß e Kruse [RK10], a RF é usada como uma técnica de regressão. Basicamente, uma RF é um método que consiste em muitas árvores de regressão e gera um resultado combinado dessas árvores como uma previsão para a variável de destino.

4. *M5-Prime* (M5P): É uma árvore de regressão que usa uma abordagem alternativa, dividindo recursivamente o espaço das amostras em pequenas regiões até que cada região seja pequena o suficiente para ser representada por um modelo simples. Gonzalez *et al.* [SAS+14b] usa árvore de regressão M5P de amostra, com características particulares para a regressão. Assim, os autores selecionaram o M5P e comparam a imprecisão com outras técnicas de regressão ao problema de previsão de rendimento das safras.

5. *K-Nearest Neighbor* (KNN): Segundo Pudumalar *et al.* [PRR+17], o algoritmo pode ser usado para classificação e regressão. É um algoritmo não complexo que armazena todos os casos disponíveis e classifica novos casos com base em alguma medida de similaridade. O conjunto de amostras é classificado com base na "proximidade" que é uma medida de distância euclidiana ou distância de Manhattan, medidas estas que medem o grau de semelhança entre as instâncias e realizam o agrupamento de acordo com a sua coesão de similaridade.

6. *Linear Regression* (LR): É uma técnica usada para analisar uma variável de resposta Y que muda com o valor da variável de intervenção X. É uma abordagem para prever o valor de uma variável de resposta a partir de um determinado valor do explicativo. Sellam e Poovammal [SP16] usam a regressão linear para analisar e determinar a relação entre a variável resposta e uma variável explicativa, tais como, a precipitação anual, área de cultivo, índice de preços de alimentos.

7. *Multilayer Perceptron* (MP): É uma rede neural artificial (RNA) de processamento simples conectada por meio de interconexões direcionadas e ponderadas. Cada unidade de processamento recebe um número de entradas do exterior ou de outras unidades de processamento. Cada entrada é calibrada com base nos pesos de suas interconexões. Uma vez calibradas, as entradas são combinadas e transmitidas para outras unidades de processamento através das interconexões apropriadas. Gonzalez *et al.* [SAS+14a] escolheram o algoritmo MLP como prioridade para a predição de rendimento das safras utilizando uma topologia de camada de 10 neurônios em uma única camada oculta, enquanto a camada de saída tem apenas um neurônio usado na estimativa da predição do rendimento das safras.

8. *Bagging*: É um método para gerar várias versões de um preditor e usa-se para obter um preditor agregado. No caso de regressão, os resultados da previsão são calculados. Ruß e Kruse [RK10] argumentam que várias versões do preditor são construídas, utilizando amostras de bootstrap do conjunto de aprendizado e usam como novos conjuntos de aprendizado. Acrescentam ainda que geralmente, o Bagging é considerado útil em configurações de regressão em que pequenas mudanças no conjunto de dados de treinamento podem causar grandes perturbações nas variáveis-alvo previstas.

9. *Additive Regression* (AR): É um classificador Meta que aprimora o desempenho de um classificador de base de regressão. Os autores Bhojani e Bhatt [BB18] usam este algoritmo de regressão para os valores contínuos de previsão, contando com um modelo de probabilidade bayesiano subjacente em vez de um algoritmo puro.

10. *Regression By Discretization* (RBD): É um esquema de regressão que emprega qualquer classificador em uma cópia dos dados que possui o atributo de classe discretizado. Geetha e Elizabeth [GS18] modificaram a regressão por discretização para

treinar e testar os conjuntos de dados a fim de obter a previsão do rendimento da cultura para o solo. O valor previsto é o esperado do valor da classe média para cada intervalo discretizado (com base nas probabilidades previstas para cada intervalo).

2.3.2.1 TÉCNICA DE AMOSTRAGEM NA AGRICULTURA

Tendo um conjunto de dados, devem ser empregados algoritmos para indução do preditor e avaliados os resultados obtidos. Para tal, devem ser utilizados métodos de amostragem alternativos para obter estimativas de desempenho preditivos mais confiáveis, definindo subconjuntos de treinamentos e de teste.

Os dados de treinamento são empregados na indução e no ajuste de modelos enquanto os de teste simulam a apresentação de objetos novos ao preditor que não foram vistos em sua indução. É utilizada na aplicação dos algoritmos preditores em conjunto de dados agrícolas a Validação Cruzada (VC) para avaliar os resultados e taxas de acerto de cada algoritmo executado.

A validação cruzada avalia a capacidade de generalização de um modelo a partir de um conjunto de dados. Segundo Xu e Goodacre [XG18], VC é amplamente empregada em problemas onde o objetivo da modelagem é a predição. Busca estimar o quão preciso é o modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

De acordo com Witten *et al.* [WFHP16], VC é um processo estatístico de partição das amostras de dados em subconjuntos onde a análise é efetuada em um conjunto inicial, enquanto outros subconjuntos são retidos para uso subsequente na validação e testes do modelo.

Existem diversas variações para o processo de VC. Na aplicação no conjunto de dados do estudo, o método de VC com k-folhas particiona o conjunto de dados original em $K = 10$ subconjuntos onde um dos subconjuntos é retido como dados de validação para testar o modelo, enquanto os 9 subconjuntos são utilizados como dados de treinamento. Este processo é repetido 10 vezes utilizando em cada ciclo uma partição diferente para o teste como é apresentada na figura 2.2.

Os 10 resultados de cada etapa podem ser combinados ou pode-se gerar a média a fim de produzir um único resultado estimado. Este é o modelo de teste mais eficaz entre as metodologias apresentadas, razão da escolha do método. Nos estudos relacionados à agricultura, a maioria dos autores aplicam a VC em seus treinamentos com k=10.

Como o problema do trabalho em estudo envolve medidas de desempenho é necessário reportar estimativas precisas de desempenho esperado pelos algoritmos e apresentar um intervalo de confiança para a média calculada que envolve também o desvio padrão obtido. De certa forma o desvio padrão auxilia na escolha entre dois algoritmos com desempenho semelhante. Razão pela qual este modelo será implementado em nosso estudo.

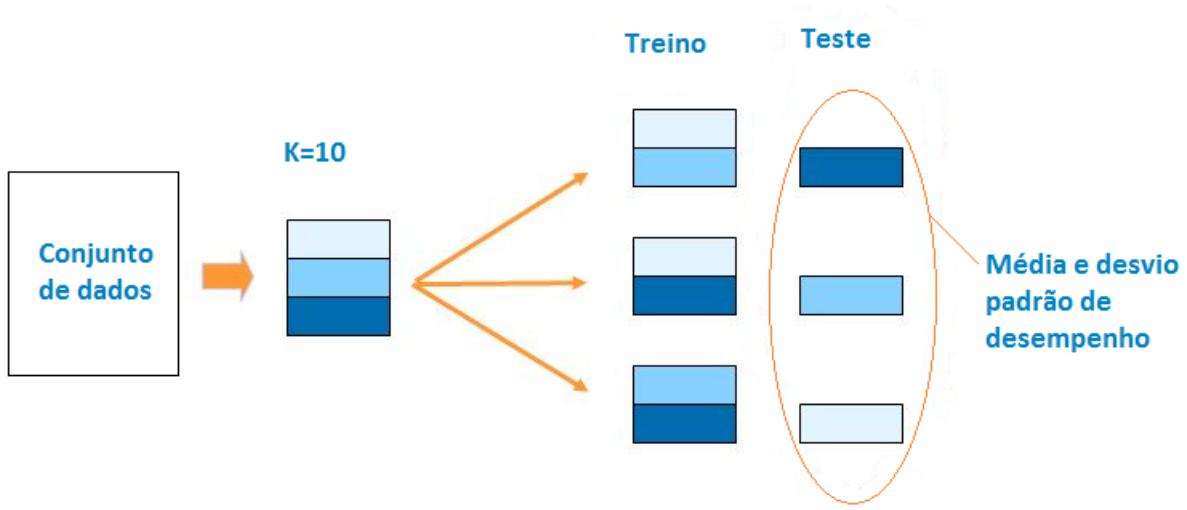


Figura 2.2: Método de validação cruzada

2.3.3 AVALIAÇÃO DOS MODELOS PREDITIVOS

De acordo com Faceli *et al.* [FLG+11] na aplicação dos algoritmos a problemas reais, em geral, o conhecimento que se tem do domínio sendo investigado é provido unicamente pelo conjunto de exemplos a partir do qual a indução de um modelo preditivo é então realizado. Depois de apresentados algoritmos preditivos que podem ser utilizados na indução de modelos de classificação e regressão na agricultura, Faceli *et al.* afirmam que não existe uma técnica universal que se sairá melhor na resolução de qualquer tipo de problema.

Há necessidade de experimentação para validar qualquer técnica proposta em que demonstre a sua efetividade na solução de diferentes problemas. Em problemas de predição, são avaliados os desempenhos dos modelos dos algoritmos obtidos nas predições realizadas. Essa avaliação de preditores é realizada por meio do desempenho do preditor gerado na rotulação de novos objetos não apresentados previamente em seu treinamento, através das medidas de erro, como Faceli *et al.* [FLG+11] argumentam em seus estudos.

Na tabela 2.2 apresentamos as métricas mais utilizadas pelos autores dos trabalhos pesquisados para a realização do estudo, em particular citamos dois usados por Gonzalez [SAS+14a] para avaliar o desempenho de modelos de regressão. Na regressão buscamos prever um valor numérico, como por exemplo, o rendimento futuro das culturas.

Nesta dissertação, utilizamos duas das métricas mais comuns dos modelos de regressão: O erro médio absoluto (MAE) e a raiz do erro médio quadrático (RMSE). O fator de correlação (R) mede a relação linear entre as previsões do modelo de regressão e os valores reais. As equações mostram como essas métricas são calculadas, onde y_i repre-

senta o valor do rendimento real, $\hat{y}$ a estimativa do rendimento e i o número da amostra. Para o R, x_i e y_i os valores das variáveis x e y e $\hat{x}$ e $\bar{y}$ são respectivamente as médias dos valores reais e previstas do rendimento x_i e y_i.

Tabela 2.2: Métricas de avaliação de modelos preditivos.

Métrica	Fórmula	Definição
Erro Médio Absoluto (MAE)	$MAE = \frac{1}{n}\sum_{t=1}^{n}\lvert y_i - \hat{y}_i)\rvert$ (2.1)	Média das diferenças absolutas entre previsões e valores reais.
Raiz do Erro Médio Quadrático (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y}_i)^2}$ (2.2)	Muito parecido com o erro absoluto médio, pois fornece uma ideia grosseira da magnitude do erro.
Coeficiente de Correlação (R)	$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$ (2.3)	Usado para medir a força de uma associação linear entre duas variáveis ou mais variáveis.

2.3.4 FERRAMENTA PARA O PROCESSO DE DCBD

De entre várias ferramentas de mineração de dados existentes, descrevemos a ferramenta WEKA por ter sido a escolhida para a execução dos algoritmos propostos no presente trabalho. A sua escolha, foi por tornar as tarefas de mineração de dados extremamente fáceis e rápidas e muito usada em pesquisas e educação. Para além de ser uma ferramenta livre, está disponível gratuitamente no site https://www.cs.waikato.ac.nz/ml/weka/ sob a Licença Pública Geral GNU.

De acordo com Jagtap *et al.* [J+13] o *Workbench* do Weka contém uma coleção de ferramentas de visualização e algoritmos para análise de dados e modelagem preditiva, junto com interfaces gráficas de usuário para facilitar o acesso a essa funcionalidade.

No WEKA, os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamados a partir do seu próprio código Java, informa Srivastava [Sri14]. A sua versão original foi projetada principalmente como uma ferramenta para analisar dados mas, a versão mais recente baseada em Java (WEKA 3) é agora usada em muitas áreas de aplicação, em particular para fins educacionais. A figura 2.3 apresenta a interface inicial do WEKA onde se seleciona a aplicação correspondente para a execução do processo. Como

exemplo, nesta interface a guia *Explorer* foi selecionada. A figura 2.4 apresenta a principal interface do WEKA onde podemos executar as etapas do processo de DCBD.

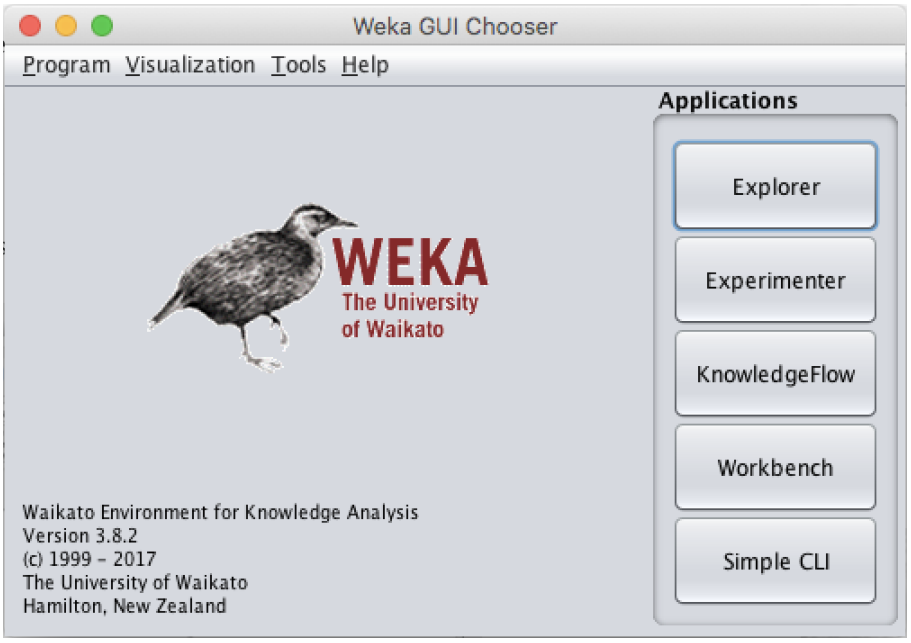


Figura 2.3: Interface Inicial da escolhas das aplicações do WEKA

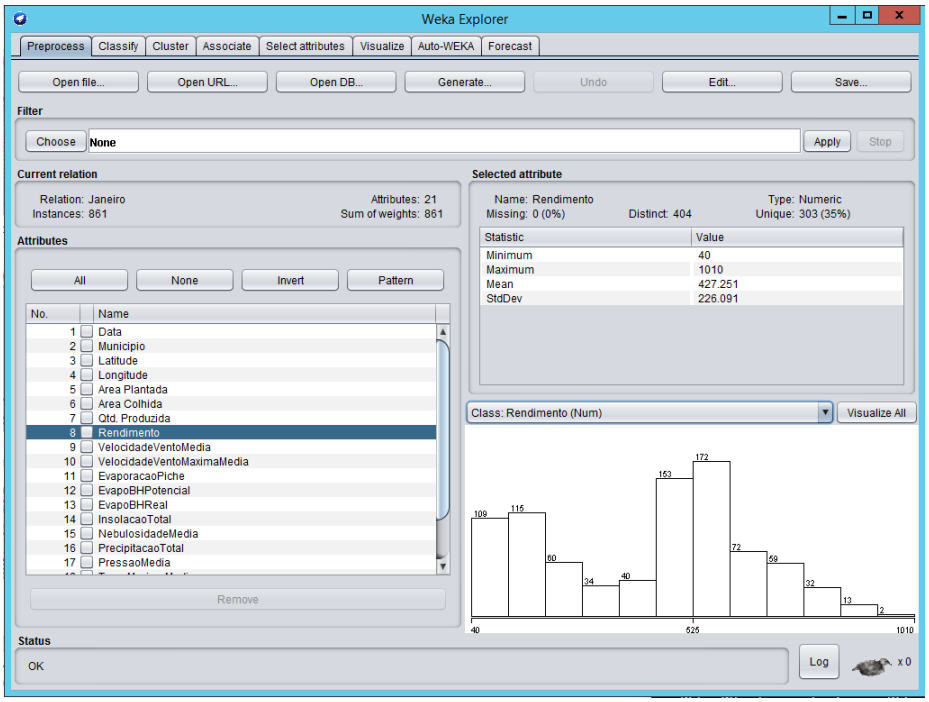


Figura 2.4: Interface do WEKA Explorer

As vantagens extraídas da ferramenta incluem:

- Disponibilidade gratuita sob a Licença Pública Geral GNU;
- Portabilidade, uma vez que é totalmente implementado na linguagem de programação Java e, portanto, é executado em quase qualquer plataforma de computação moderna;
- Possui uma coleção abrangente de pré-processamento de dados e técnicas de modelagem;
- Facilidade no uso devido às interfaces gráficas do usuário Weka que suporta várias tarefas de mineração de dados padrão, mais especificamente pré-processamento, agrupamento, classificação, regressão, visualização e seleção de recursos (https://www.cs.waikato.ac.nz/ml/weka/).

Todos os algoritmos executados pela ferramenta baseiam-se na suposição de que os dados estão disponíveis como um único arquivo simples. Cada dado é descrito por um número fixo de atributos, normalmente, atributos numéricos ou nominais. No entanto, outros tipos de atributos também são suportados, acrescenta Srivastava [Sri14].

3. MÉTODO

Apresentamos nesta sessão o método de DCBD reutilizável. Neste método serão definidas em etapas, a composição da estrutura básica proposta para a realização da reutilização de um processo de DCBD na agricultura.

O método de reutilização é aplicado apenas à culturas temporárias devido ao tipo de informação que a cultura apresenta. A natureza diversificada de culturas pode gerar modelos diferentes na execução do método, o que implica dificuldades na obtenção e construção do método ideal capaz de resolver tal situação. Para compensar essas lacunas, busca-se trabalhar no presente estudo, dentro do mesmo domínio onde encontramos características semelhantes no conjunto de dados como a periodicidade de plantação, desenvolvimento e colheita.

Para além da agilidade que o método prevê, faz sugestões dos algoritmos com melhor desempenho para a previsão do rendimento das culturas que podem ajudar o setor agrícola na tomada de decisão no planejamento das atividades agrícolas, aumentando assim, a qualidade das culturas e o número de culturas previstas em um curto período de tempo.

3.1 ESTRUTURA DO MÉTODO

A figura 3.1 ilustra a estrutura do método de reutilização do processo de DCBD proposta. De acordo com a figura 3.1, é composta por três etapas:

1. Etapa do processo de DCBD: onde são determinadas as tarefas a serem reutilizadas em novos processos;
2. Etapa da base de conhecimento: onde são catalogadas as tarefas executadas na etapa do processo de DCBD;
3. Etapa da reutilização: onde o processo DCBD é executado a partir do uso da base de conhecimento que apoia a sua execução.

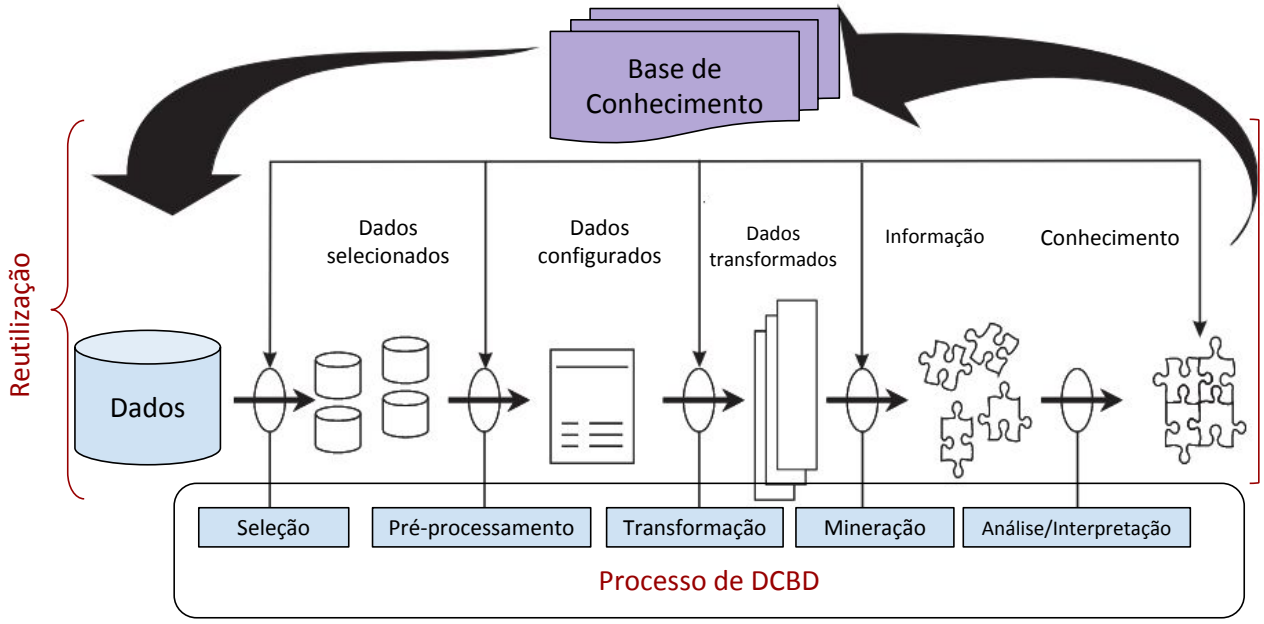


Figura 3.1: Método da reutilização do processo de DCBD agrícola.

A primeira etapa do método, etapa do **processo de DCBD**, segue a metodologia proposta por Fayyad, Piatetsky-Shapiro e Smyth na sua execução. A primeira fase do processo, seleção, segundo Manjundar *et al.* [MNA17] é a fase que busca obter e avaliar dados entre diversas fontes de dados existentes formando assim um conjunto de dados usados no processo de DCBD. É nesta fase que dados selecionamos dados que vão compor o conjunto de dados usados no presente trabalho.

O pré-processamento é a segunda fase do método que avalia a integridade das informações disponibilizadas (Manjundar *et al.* [MNA17]). Nesta fase são manipulados os dados selecionados através de aplicação de técnicas para mapeamento, ajustes, tratamento, limpeza dos dados procurando adequar para a obtenção dos resultados definidos nos objetivos do processo.

Segundo Azevedo [Aze19], na fase de transformação são aplicadas técnicas de tratamento de dados responsáveis pela padronização dos dados. A padronização é uma tarefa essencial para o processo pois, em função da natureza dos dados, procura-se padronizar para que não apareçam com formatos diferentes e melhorando assim os resultados apresentados pelos algoritmos de mineração.

A fase de mineração é principal fase de todo o processo, informam Manjundar *et al.* [MNA17]. Segundo os mesmos, é onde o conhecimento existente nos dados é extraído. Nesta fase são executados algoritmos responsáveis por encontrar os padrões contidos nos dados selecionados afim de gerar informação relevante para o estudo.

A fase de avaliação/interpretação dos resultados determina a eficácia do desempenho dos métodos através de testes executados na fase de mineração. Nesta etapa são

efetuados testes com diversos parâmetros que cada algoritmo possui, visando alcançar a configuração mais próxima do resultado ótimo para o conjunto de dados da agricultura.

Na etapa da **Base de Conhecimento** são catalogadas todas as configurações necessárias executadas no processo de DCBD agrícola sobre uma determinada cultura. Este conhecimento apreendido da execução base será reutilizado para a execução de um novo conjunto de dados agrícolas do mesmo domínio.

Esta etapa vai disponibilizar essas tarefas configuradas em forma documental e de fácil uso pelos usuários. Vai permitir que um usuário comum e não especifico do domínio do processo, realize o processo com facilidade.

A etapa da **Reutilização** é a etapa onde realiza-se o processo de DCBD usando as catalogações contidas na base de conhecimento. Ele permitirá a realização do processo sem entrar em detalhes das analises para descobrir quais atividades devem ser realizadas no processo de DCBD.

3.2 EXECUÇÃO DO MÉTODO

Executamos a seguir, o método, descrevendo as três etapas listadas. No entanto, antes de se iniciar com a execução das etapas do método é necessário, entender o problema que estabelece qual o objetivo que se deseja atingir com o processo de DCBD no setor agrícola, entender o cenário no qual estamos inseridos e estudar a área da seleção da cultura.

1. Estudo da área da cultura

   O Estado do Rio Grande do Sul (RS) foi o local escolhido para execução do processo sobre a pesquisa. Ele foi escolhido por ser o maior produtor de arroz com 25,6% da área cultivada e 44,5% da produção segundo a Pesquisa Agrícola Municipal do IBGE [dGeEI19].

   Segundo o Atlas Sócio-Econômico do RS [dRGdS19], localiza-se entre os paralelos 27° 03"42" " e 33° 45" 09" " de latitude Sul, e 49° 42" 41" " e 57° 40" 57" " de longitude Oeste, com cerca de 11,29 milhões de habitantes, do último censo de 2016 e possui um total de 497 municípios.

   As temperaturas apresentam grande variação sazonal, com verões quentes e invernos bastante rigorosos, com a ocorrência de geadas e precipitação eventual de neve. As temperaturas médias variam entre 15℃ e 18℃, com mínimas de até -10℃ e máximas de 40℃. Com relação às precipitações, o Estado apresenta uma distribuição relativamente equilibrada das chuvas ao longo de todo o ano, em decorrência das massas de ar que penetram no Estado [dRGdS19].

2. Estudo do impacto das mudanças climáticas na agricultura

   Segundo Amanda Balbino [Bal19] a agricultura é uma atividade altamente dependente de fatores climáticos. Por isso, a mudança no clima pode afetar a produção agrícola de várias formas: mudança na severidade de eventos extremos, no número de graus-dia de crescimento, devido as alterações na temperatura do ar, modificação na ocorrência e na severidade de pragas e doenças, dentre outros.

   Os responsáveis dos setores agrícolas precisam entender melhor o ambiente onde vão plantar para tomar melhores decisões. É fundamental que estes incorporem a gestão de riscos decorrentes das mudanças climáticas em seus processos de planejamento. Parte disto está associado à identificação da vulnerabilidade atual aos impactos de eventos climáticos [Bal19].

3. Definição do objetivo do processo

   Este processo tem como objetivo prever o rendimento das culturas por forma a ajudar o setor agrícola na tomada de decisão certa.

O estudo descrito acima serve como base para a seleção dos dados que vão constituir o conjunto de dados a serem usados no processo de DCBD aplicado no setor agrícola.

3.2.1 PROCESSO DE DCBD

Este processo segue a metodologia definida pelos autores Fayyad, Piatetsky-Shapiro e Smyth, incorporando as cinco fases do processo. Neste método, ele é executado inicialmente para identificação das tarefas, ajustes e configuração que os dados selecionados podem necessitar para posterior reaproveitar em novos processos.

Embora cada fase do processo de KDD seja independente, podendo ser tratada individualmente, existe uma forte dependência entre elas. Assim, para que seja feita uma correta configuração dos dados, é necessário ter uma base de dados corretamente modelada. No entanto, segue-se a execução das cinco fases do processo de DCBD.

3.2.1.1 SELEÇÃO DOS DADOS

Foi definido um período em série temporal para trabalhar o processo de DCBD que é de 5 anos (2013 a 2017). Verificou-se que o clima é importante para a produção agrícola e por causa disso, os seus dados são necessários. Para complementar estes dados, precisou-se das regiões em um menor grau possível para um bom resultado do processo relacionando com as coordenadas geográficas.

Selecionou-se os conjuntos de dados com a descrição na tabela 3.1 para a serem executados no processo de DCBD. Os dados selecionados, foram obtidos de diferentes fontes de dados, apresentados:

- Do Levantamento Sistemático de Produção Agrícola (LSPA) do Instituto Brasileiro de Geografia e Estatística (IBGE) [dGeE18], buscou-se dados de Produção agrícola, como a figura 3.2 apresenta em seu exemplo.

Municipio	2013	2014	2015	2016	2017
Alta Floresta D'Oe	480	336	336	336	235
Ariquemes (RO)	1.600	1.600	1.360	1.160	812
Cabixi (RO)	3.480	4.000	2.765	1.860	2.230
Costa Marques (R	60	30	30	-	-
Espigão D'Oeste (	750	850	255	55	35
Guajará-Mirim (R	100	36	12	12	12
Jaru (RO)	123	5	405	8	8
Ji-Paraná (RO)	20	20	10	-	-

Figura 3.2: Conjunto de dados de produção agrícola de Arroz

- Do Banco de Dados de Pesquisas Meteorológicas (BDMET) [Met18], buscou-se os dados climáticos, como a figura 3.3 apresenta em seu exemplo.

Estacao	Municipio	Data	Velocidade	Evaporaca	EvapoBHPot	EvapoBHRea
83980	Bage(RS)	31/01/2013	28	158.5	119.16	119.16
83980	Bage(RS)	30/11/2013	33	126	94.6	94.6
83980	Bage(RS)	31/12/2013	28	212.4	141.46	104.82
83980	Bage(RS)	31/01/2014	28	156.6	144.87	144.87
83980	Bage(RS)	28/02/2014	29	116.6	11.3	11.3
83980	Bage(RS)	31/03/2014	24	98.2	84.8	84.80

Figura 3.3: Conjunto de dados Meteorológicos

- Do Departamento de Astronomia do Instituto de Física da Universidade Federal do Rio Grande do Sul (DA/IF-UFRGS), buscou-se dados das coordenadas geográficas [DA-18] relacionada a cada município de produção do arroz, como a figura 3.4 apresenta em seu exemplo.

Data	Municipio	Latitude	Longitude
nov/13	Hulha Negra	-31.4067	-53.8667
jul/14	Hulha Negra	-31.4067	-53.8667
dez/15	Hulha Negra	-31.4067	-53.8667
abr/16	Hulha Negra	-31.4067	-53.8667
jan/17	Hulha Negra	-31.4067	-53.8667

Figura 3.4: Conjunto de dados das coordenadas geográficas

Tabela 3.1: Descrição dos dados

Dados	Descrição
Período de seleção	2013 a 2017
Municípios do RS	236 municípios produtores do arroz.
Dados Agrícolas	Área colhida (hectares), Área plantada (hectares), Quantidade produzida (toneladas/hectares) e Rendimento da cultura (toneladas/hectares).
Dados das Coordenadas Geográficas	Latitude e Longitude (decimais).
Dados Meteorológicos	Velocidade do vento (média e máxima), Evaporação do ar (Piche, BHreal e BHpotencial), Insolação Total, Nebulosidade média, Precipitação Total, Pressão (Média e Nível Mar Média), Temperatura (Média, Compensada, mínima) e Umidade Relativa Média.

Durante a seleção dos dados, na busca pelos dados agrícolas, verificou-se que as fontes disponibilizaram dois conjuntos de dados com formatação diferente: dados agrícolas com totais anuais de produção para cada município produtor do RS e dados agrícolas com totais mensais de produção para todo o Estado do RS. Selecionou-se o conjunto de dados com a lista dos municípios do RS que produzem o arroz para facilitar o processo de previsão.

3.2.1.2 PRÉ-PROCESSAMENTO

Os conjuntos de dados selecionados apresentam características, dimensões e formatos diferentes. No entanto, as técnicas de pré-processamento têm por objetivo melhorar a qualidade dos dados.

Os dados adquiridos das fontes, sobre quais serão pré-processados, foram incorporados na planilha do Microsoft Office Excel para melhor exploração dos dados colhidos. Cada conjunto de dados foi analisado em separado na respectiva planilha para verificar os atributos que cada conjunto de dados possui.

3.2.1.2.1 AJUSTE NO CONJUNTO DOS DADOS

Na previsão do rendimento da cultura deve-se analisar cuidadosamente os dados de produção selecionados. Tratando-se de culturas temporárias, a informação mensal dos dados é pertinente, pois culturas temporárias, segundo Carmo, [Car15] são aquelas sujeitas ao replantio após a colheita, ou seja, que devem ser plantadas a todo ano, após a colheita, e geralmente em um curto período de tempo.

Como os dados selecionados da cultura são anuais, como pode-se ver na figura 3.2, o não condiz com o formato ideal, deve-se ajustar os dados em mensais. Assim o

método poderá prever corretamente o rendimento da cultura. O ajuste foi feito na base do cálculo para a distribuição do valor anual de cada município em valores mensais. Foi feita uma distribuição percentual mensal a cada município, no período correspondente, usando a regra de três simples dado pela fórmula:

$$Vmês = \frac{total\_anual \times percentagem\_mensal}{100} \tag{3.1}$$

Em que:

- *Vmês*: é o valor mensal procurado correspondente a cada município;
- *total_anual*: é o valor total anual existente do município correspondente;
- *percentagem_mensal*: é a percentagem mensal atribuída a cada município por ano.

A planilha do Excel foi usada para o cálculo do ajuste. Nela foram anotadas as operações dos ajustes dos percentuais usados que poderão ser reutilizados em uma outra execução. No final, os dados agrícolas apresentaram o formato ajustado para culturas temporárias, contendo informação mensal de janeiro a dezembro no período de 2013 a 2017 conforme a figura 3.5 apresenta. Os outros dados colhidos apresentavam informação mensal apropriada e não necessitaram de nenhuma configuração.

Municipio	jan/13	mar/14	jun/15	set/16	dez/17
Alta Floresta DOe	38.4	20.496	24.528	30.576	14.1
Ariquemes (RO)	128	97.6	99.28	105.56	48.72
Cabixi (RO)	278.4	244	201.845	169.26	133.8
Cacoal (RO)	44.24	16.836	20.148	25.116	3
Cerejeiras (RO)	104	79.3	94.9	100.1	60
Colorado do Oest	12	9.15	2.555	6.37	4.2
Corumbiara (RO)	640	183	219	273	120
Costa Marques (R	4.8	1.83	2.19	-	-
Espigão DOeste (F	60	51.85	18.615	5.005	2.1

Figura 3.5: Conjunto de dados mensal da cultura de arroz

3.2.1.2.2 INTEGRAÇÃO DOS CONJUNTOS DOS DADOS

No processo de DCBD, os dados devem estar no mesmo conjunto de dados. Isto permite que o conhecimento esperado seja encontrado a partir de avaliação conjunta das variáveis contidas no conjunto de dados. Sendo que os dados selecionados vem de fontes diferentes, por essa razão, são encontrados em conjunto de dados separados. Este problema é comum quando se busca resolver problemas em que se necessita de diferentes informações para compor o estudo.

A solução para este problema é integrando os dados selecionados a um único conjunto de dados. Segundo Lenzerini, [Len02], o processo de integração de bases de

dados, envolve a combinação de dados que residem em diferentes fontes e fornecem uma visão unificada dos dados e os torna significativos em diversas situações. A figura 3.6 ilustra como o processo de integração de dados é construído sobre os dados selecionados.

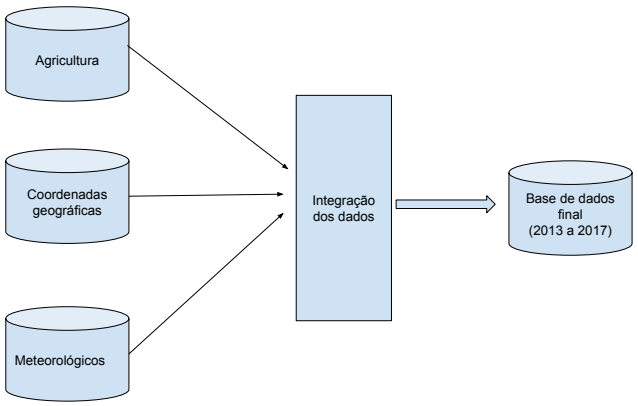


Figura 3.6: Processo de integração de dados. Adaptado do Lenzerini [Len02].

Usou-se a planilha do Excel para integrar os dados em um único conjunto de dados. A integração é feita unificando os conjuntos de dados através de variável comum existente nos conjuntos de dados. No caso a variável "Município" foi a variável da correspondência usada para a unificação dos conjunto onde também foi respeitado o período correspondente de cada variável no processo de integração. Após este processo de integração o conjunto de dados da cultura do arroz, o conjunto de dados selecionados tornou-se um único conjunto contendo toda a informação necessária dos conjuntos de dados selecionados, como é apresentada na figura 3.7.

Data	Municipio	Latitude	Longitude	Area Plan	Area Colh	Qtd. Prod	Rendimer	VelocidadeVento	VelocidadeVen	Evaporaca
nov/13	Hulha Negra	-31.4067	-53.8667	110	110	110	725	24	14	104.5
jul/14	Hulha Negra	-31.4067	-53.8667	91.3	91.3	91.3	601.75	24	14	104.5
dez/15	Hulha Negra	-31.4067	-53.8667	81.6	81.6	81.6	510	24	14	104.5
abr/16	Hulha Negra	-31.4067	-53.8667	118.56	118.56	118.56	576	24	14	104.5
jan/17	Hulha Negra	-31.4067	-53.8667	759.78	759.78	759.78	673.758	24	14	104.5

Figura 3.7: Conjunto de dados final da cultura de arroz

3.2.1.2.3 TRATAMENTO E LIMPEZA DOS DADOS

No conjunto de dados formado, as variáveis ou atributos apresentavam formatação e características de escritas diferentes, em variáveis semelhantes, que foram padronizadas ao mesmo método. Alguns atributos apresentavam caracteres que não eram suportadas pela ferramenta WEKA, em uso no processo, estes foram removidos e outros foram adequados para uma melhor compreensão do especialista ou utilizador. Usou-se a planilha do Excel para identificar de forma fácil e rápida esses caracteres e permitir a sua substituição automática sem alterar as outras informações contidas no conjunto de dados.

Durante o processo de integração dos dados, observou-se que no conjunto de dados meteorológicos, havia poucos municípios com estações meteorológicas o que culminou com a integração para facilitar a análise e interpretação, dada a falta de dados meteorológicos de municípios com produção agrícola do arroz.

A falta de dados em um processo de DCBD pode interferir no resultado do objetivo do processo, por tanto, é necessário trata-los com a interpolação.

Existem diversas formas de tratamento de dados faltantes dependendo da formatação que estes apresentam para a sua aplicação. Para o nosso caso em estudo, uma vez que estes dados são desconhecidos ou não existem pela falta de estações meteorológicas, pode ser resolvido através da interpolação dos dados a partir dos dados conhecidos das estações.

Foi calculado o valor a interpolar, pela fórmula das médias absolutas usando valores dos municípios das estações vizinhas. Calculou-se primeiramente, a distância latitudinal a partir dos valores das coordenadas geográficas de cada município para encontrar os municípios que se avizinham, usando a fórmula definida por Delfino [Del19]:

$$DLA = |LA_f - LA_i| \tag{3.2}$$

Onde:

- DLA é a distância latitudinal,
- LA_f é latitude final que refere-se a um município e,
- LA_i é latitude inicial, a um outro ponto de um município.

Conhecendo os municípios vizinhos, calculamos as médias absolutas dos valores como sendo os valores a serem interpolados nos municípios com falta de dados meteorológicos dada pela equação:

$$Valor\_inter = \frac{soma\_valores * total\_observacoes}{total\_observacoes} \tag{3.3}$$

Em que:

- *Valor_inter*: é o valor da interpolação procurado;
- *soma_valores*: é o somatório dos valores das estações mais próximas;
- *total_observaes*: é o total dos municípios a serem interpolados.

Toda esta operação foi realizada na planilha do Excel para cada município produtor da área do RS, onde procuramos guardar formatação dos cálculos para reutilizar em outros casos.

Para o caso em que a falta de dados surgiu devido a dados não conhecidos ou não informados no conjunto de dados, foi imputado o caractere "?" no seu preenchimento desses campos. Este caractere é reconhecido pela ferramenta de mineração de dados, WEKA, como um valor faltante que não interfere nos resultados obtidos dos algoritmos de mineração de dados.

3.2.1.2.4 ANÁLISE DO CONJUNTO DE DADOS

Aparentemente o conjunto de dados encontra-se na forma ideal para a mineração de dados, sendo que é de grande importância fazer uma análise no conjunto de dados para entender a correlação existente entre as variáveis do conjunto, sendo que ele surgiu da integração de diferentes conjuntos para compor um único conjunto com todas as variáveis necessárias para o processo.

Analisar a existência da correlação entre as variáveis do conjunto de dados, ajuda a entender a sua influência na agricultura ao prever o rendimento das culturas. Essa influência pode apresentar a correlação das variáveis de forma independente ou dependente uma da outra.

Correlacionar variáveis segundo Faul *et al.* [FEBL09] é medir e descrever a relação entre duas variáveis numéricas x e y através da representação dos pontos para todas as variáveis que representam o conjunto de dados. Assim, obtém uma dispersão de pontos que podem sugerir uma relação entre as variáveis. Essa dispersão é feita através de uma análise estatística que ajuda na compreensão das variáveis.

O coeficiente das correlações (R) pode variar entre -1 e $+1$ medindo assim, o grau de correlação sendo ela forte negativa ou forte positiva, representada a vermelho no gráfico 3.8 e a azul a fraca negativa e forte negativa como podemos observar.

Assim sendo, realizamos a análise de correlação sobre o conjunto de dados da cultura do arroz e verificou-se que existe relação entre as variáveis selecionadas para a execução do processo. O gráfico 3.9 das correlações obtidas, as bolinhas a azul representam as correlações positivas e as vermelhas as correlações negativas variando a tonalidade para medir o nível de relação existente entre as variáveis, como demostradas no gráfico 3.8.

[1] Extraído do https://operdata.com.br/blog/coeficientes-de-correlacao/

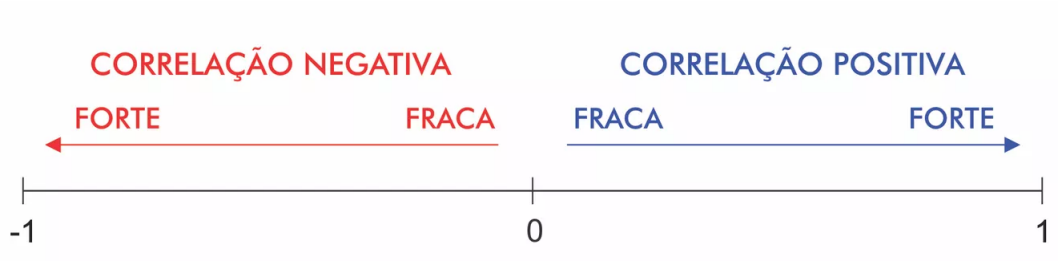


Figura 3.8: Gráfico das correlações.[1]

Ao analisar-se a variável do rendimento como exemplo, observamos que ela possui uma relação positiva com a temperatura mínima, evaporação, nebulosidade e a precipitação para dados meteorológicas e com todas as variáveis agrícolas e também relaciona negativamente com umidade, temperatura máxima, velocidades do vento, evaporação e pressão do ar. A temperatura compensada e evaporação, o gráfico 3.9 mostra que essas variáveis relacionam-se com a variável do rendimento de forma independente.

Este relacionamento entre as variáveis do conjunto ajudará na melhora do processo através do conhecimento do relacionamento entre as variáveis, possibilitando a obtenção do modelo que melhor se ajuste às variáveis que compõem do conjunto.

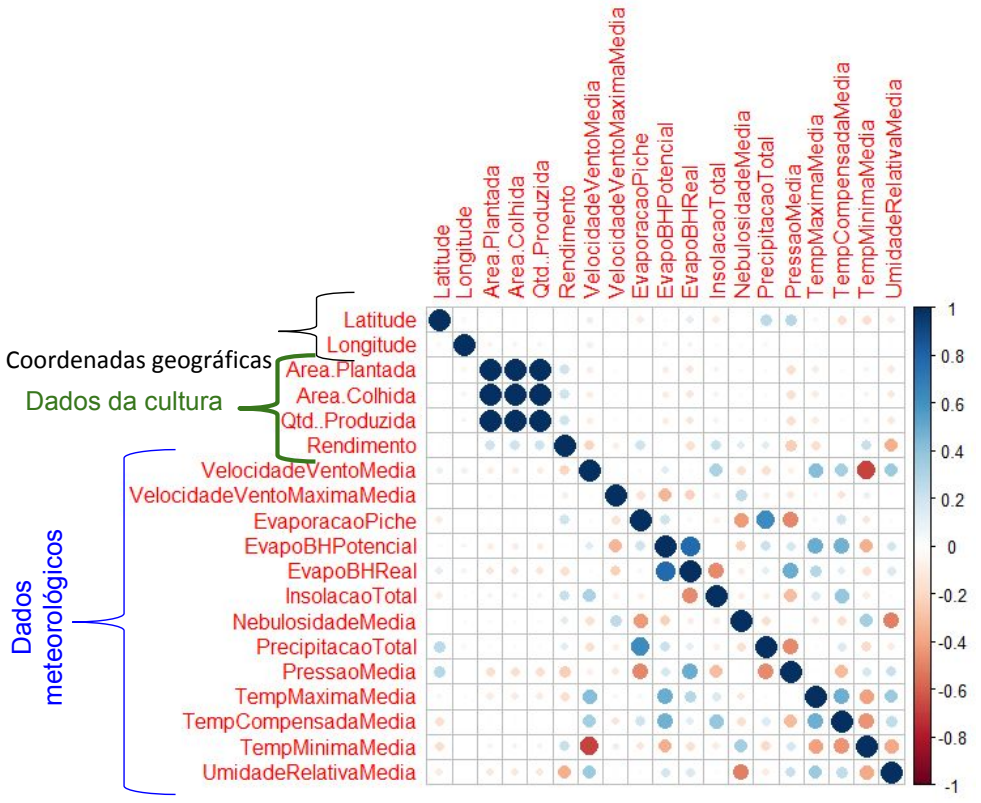


Figura 3.9: Gráfico das correlações entre as variáveis da cultura do arroz

A construção dos gráficos das correlações usadas neste estudo, foi feita com recurso ao software Rattle. Rattle é uma interface gráfica do usuário para mineração de dados que usa a linguagem R. Esta ferramenta se encontra gratuitamente disponível no endereço https://rattle.togaware.com/.

Existindo uma relação entre as variáveis selecionadas e principalmente com a variável alvo do processo de DCBD, rendimento das culturas, que nos guia ao objetivo do processo, prever o rendimento das culturas, procuramos analisar o comportamento do rendimento ao longo do período selecionado do estudo.

Através do gráfico construído do rendimento, procuramos demostrar o comportamento do rendimento ao longo dos anos de 2013 a 2017. Nele podemos observar que, por exemplo, em janeiro de 2013 o rendimento foi baixo comparando com os meses de janeiro de 2014 a 2017. Também podemos verificar uma variação em relação a cada ano sendo que o ano de 2013 vários meses apresentaram rendimento muito baixo e os anos 2014, 2016 e 2017 foram mais estáveis embora uma variação em relação aos meses do respectivo ano.

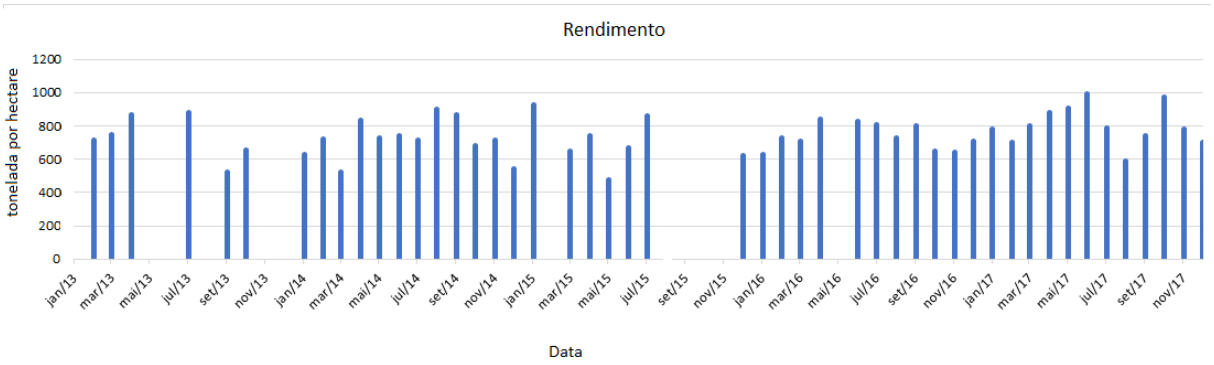


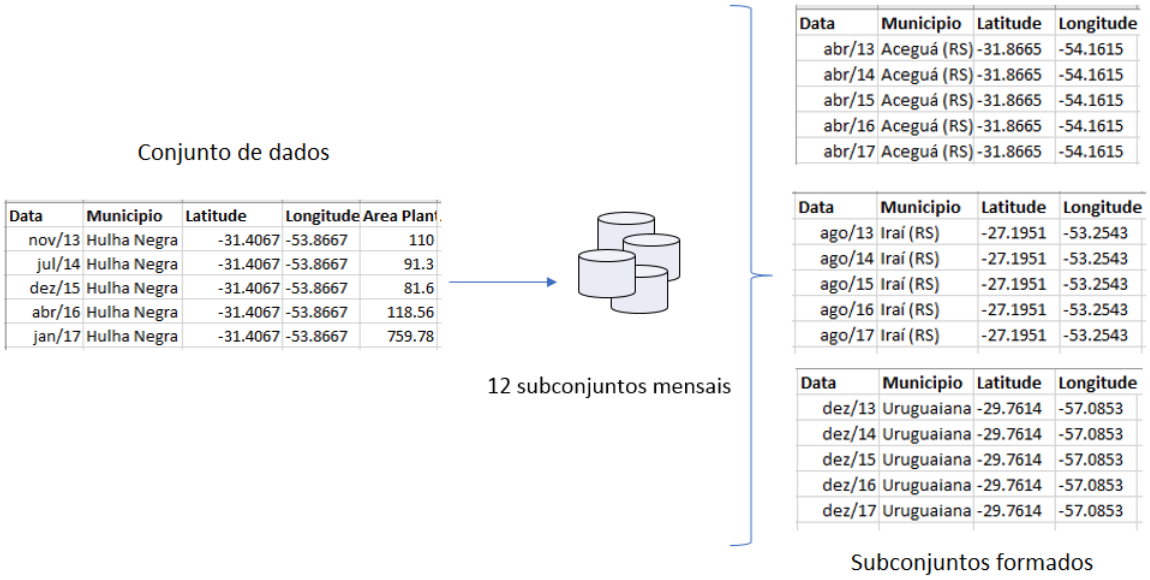
Figura 3.10: Gráfico comportamental do rendimento do arroz

Tratando-se de culturas temporárias, uma forma ideal é prever o rendimento de mês e de um ano respectivamente. Seguindo este raciocínio, agrupamos os meses independente do ano e estratifica-los para uma análise mais clarificada do rendimento. Procedendo dessa forma, podemos compreender melhor o motivo da grande diferença do rendimento entre os meses do ano.

3.2.1.2.5 ESTRATIFICAÇÃO DO CONJUNTO DE DADOS

A estratificação surge como alternativa para análise comportamental mensal do conjunto de dados. Uma vez que estamos perante a execução do processo sobre um domínio temporário em que se trata de culturas de replantio, estes possuem períodos curtos entre a plantação e a colheita o que pode ter contribuído para o varição do comportamento que o conjunto de dados apresentou.

O processo de estratificação poderá melhorar o resultado da previsão mensal do rendimento da cultura através de uma análise realizada a cada conjunto de dados estratificado. Segundo Cano *et al.* [CHL06] o processo de estratificação de base de dados consiste em dividir o conjunto de dados em subgrupos. Assim, o conjunto foi estratificado em 12 meses do ano sem levar em consideração os anos colhidos. A figura 3.11 apresenta uma ideia do processo de estratificação ocorrida no conjunto de dados geral.



Data	Municipio	Latitude	Longitude	Area Plant
nov/13	Hulha Negra	-31.4067	-53.8667	110
jul/14	Hulha Negra	-31.4067	-53.8667	91.3
dez/15	Hulha Negra	-31.4067	-53.8667	81.6
abr/16	Hulha Negra	-31.4067	-53.8667	118.56
jan/17	Hulha Negra	-31.4067	-53.8667	759.78

Data	Municipio	Latitude	Longitude
abr/13	Aceguá (RS)	-31.8665	-54.1615
abr/14	Aceguá (RS)	-31.8665	-54.1615
abr/15	Aceguá (RS)	-31.8665	-54.1615
abr/16	Aceguá (RS)	-31.8665	-54.1615
abr/17	Aceguá (RS)	-31.8665	-54.1615

Data	Municipio	Latitude	Longitude
ago/13	Iraí (RS)	-27.1951	-53.2543
ago/14	Iraí (RS)	-27.1951	-53.2543
ago/15	Iraí (RS)	-27.1951	-53.2543
ago/16	Iraí (RS)	-27.1951	-53.2543
ago/17	Iraí (RS)	-27.1951	-53.2543

Data	Municipio	Latitude	Longitude
dez/13	Uruguaiana	-29.7614	-57.0853
dez/14	Uruguaiana	-29.7614	-57.0853
dez/15	Uruguaiana	-29.7614	-57.0853
dez/16	Uruguaiana	-29.7614	-57.0853
dez/17	Uruguaiana	-29.7614	-57.0853

Figura 3.11: Estratificação do conjunto de dados geral

Os extratos obtidos foram definidos de forma que em cada grupo de amostragem fossem semelhantes uma vez que a variabilidade minimizada dos extratos melhora a precisão das estimativas, o que torna a técnica de amostragem mais eficiente quando executados os algoritmos de mineração de dados [CHL06].

Tendo os novos conjuntos de dados, analisou-se cada um deles, para averiguar se os resultados das relações entres variáveis dos conjuntos permanecem correlacionadas e se existe alguma diferença em relação a cada mês do ano. Dos gráficos mensais das correlações apresentadas, podemos observar que existe correlação positiva e negativa variando o valor de forte para fraco representado a azul e vermelho nas bolinhas.

O único diferencial é que podemos acompanhar nesta análise, o comportamento para cada mês pois eles apresentam diferenças, por exemplo, no gráfico 3.12 do mês de janeiro o rendimento apresenta uma correlação positiva em relação às variáveis agrícolas e em relação à temperatura mínima, umidade, insolação, evaporação do ar, nebulosidade e a precipitação enquanto que no gráfico 3.13 do mês de fevereiro, o rendimento correlaciona-se positivamente com a temperatura miníma, insolação, evaporação do ar, temperatura compensada, nebulosidade e precipitação para além das variáveis agrícolas. Podemos também, verificar essas diferença nos restantes conjuntos de dados.

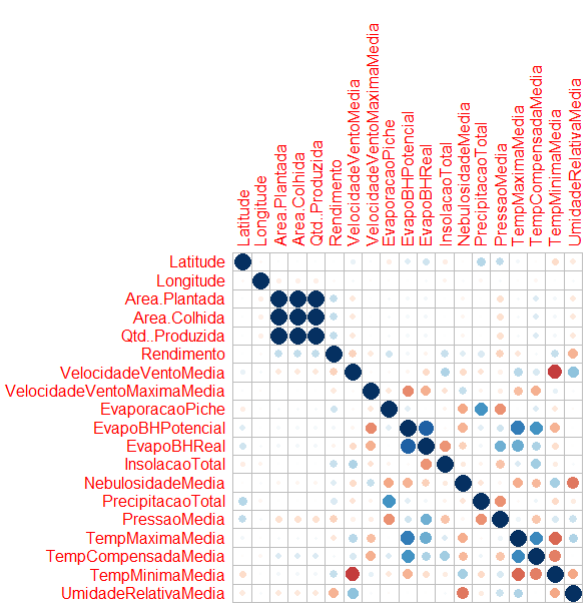


Figura 3.12: Correlação de janeiro

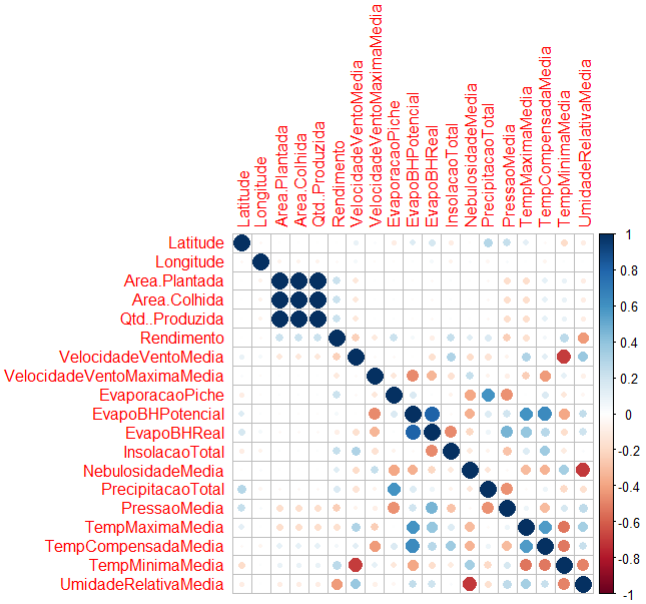


Figura 3.13: Correlação de fevereiro

Na análise do comportamento do rendimento nos doze subconjuntos apresentados nas figuras 3.14 para o mês de janeiro e na figura 3.15 para o mês de fevereiro, tornou-se mais visível o ano em que houve um baixo rendimento. Também pode-se verificar, de forma mais clara, a variação em relação ao período selecionado.

No gráfico do mês de janeiro, do ano de 2013 o rendimento foi quase que inexistente e o mesmo acontece para o rendimento de fevereiro do ano de 2015. Também podemos verificar a variação entre cada ano sendo que janeiro de 2015 o rendimento foi mais alto para o mês de janeiro e em fevereiro deu-se um elevado rendimento no ano de 2016.

Existem vários motivos para o resultado que os gráficos apresentam, sendo que isto pode estar ligado a variação climática ao longo do período que contribuiu negativamente para a produção da cultura do arroz e consequentemente o seu rendimento. Sendo que o clima contribui tanto que positivo assim como negativamente na agricultura dependendo do período de plantio, desenvolvimento e colheita.

Outro motivo pode ser pela falta de informação/dados em relação ao rendimento dentro do período selecionado. A questão da falta de dados apresentados no gráfico são considerados dados faltantes ou desconhecidos e podem surgir devido a uma falha durante o lançamento desses dados no conjunto ou pela falta de produção nesse período, entre outros.

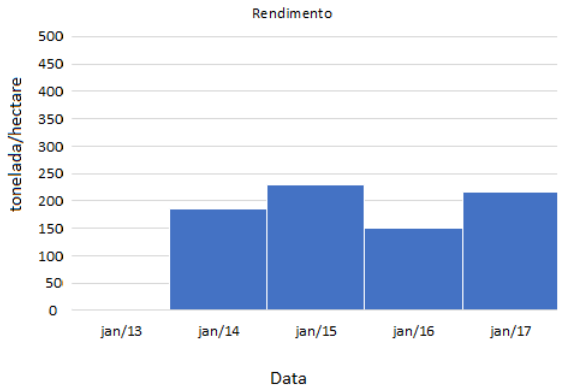


Figura 3.14: Rendimento de janeiro

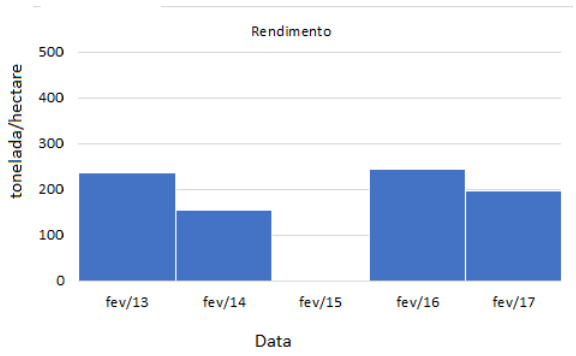


Figura 3.15: Rendimento de fevereiro

3.2.1.3 TRANSFORMAÇÃO

Uma vez verificado que a melhor estratégia para o nosso problema é fazer uma previsão mensal dos dados estratificados, passou-se a trabalhar com os doze subconjuntos de dados analisados de forma independente. Como os dados dos conjuntos vêm de fontes diferentes, eles aparecem com diferentes formatos numéricos e com padrões diferentes. No processo de DCBD os dados devem estar padronizados e devem apresentar uma forma adequada para a execução dos algoritmos, gerando assim um bom resultado.

A Normalização dos dados é uma transformação que tem a função de deixar essas variáveis de forma adequada e ela envolve a aplicação de equações matemáticas para atingir o seu objetivo.

Como os algoritmos classificadores utilizados nesse estudo exigem atributos contínuos no conjunto de dados de treinamento, optou-se em realizar o processo de normalização dos atributos quantitativos. Segundo Han e Kamber [HPK11] a normalização consiste em converter os valores de atributos para faixas de –1 a 1 ou de 0 a 1, sendo de grande utilidade para algoritmos de mineração de dados.

Usou-se a ferramenta WEKA para transformar os dados dos subconjuntos aplicando o filtro de normalização que possui o filtro *Normalize*. O filtro realiza transformação linear dos dados, considerando min_v e max_v os valores mínimo e máximo de um atributo V, e que um novo valor new_v de V é mapeado para a faixa [novo min_v, novo max_v], a partir da fórmula:

$$new_V = \frac{V - min_V}{max_V - min_V} \tag{3.4}$$

Fez-se testes preliminares com os algoritmos preditores utilizados no estudo para verificar o ganho de eficiência no desempenho dos algoritmos com aplicação do filtro, figura 3.16 e sem aplicação do filtro, figura 3.18. O resultados mostram que a aplicação do filtro nos conjuntos de dados tornam os resultados mais eficientes, como podemos ver na figura 3.17 com conjunto normalizados em relação ao conjunto de dados não normalizados apresentados na figura 3.19.

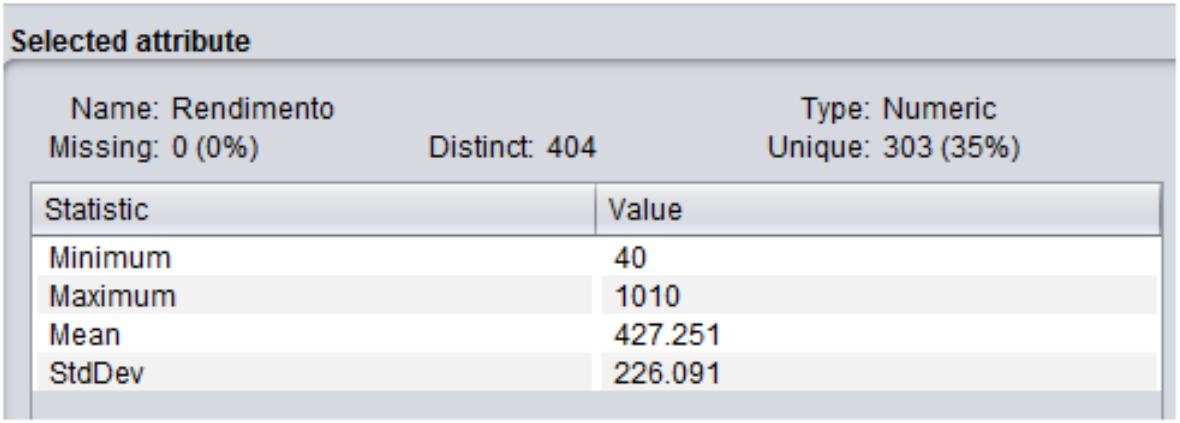


Figura 3.16: Conjunto de dados não normalizados

```
Time taken to build model: 1.61 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9369
Mean absolute error                     51.1998
Root mean squared error                 79.0447
Relative absolute error                 26.0324 %
Root relative squared error             34.9501 %
Total Number of Instances              861
```

Figura 3.17: Resultado do conjunto não normalizados

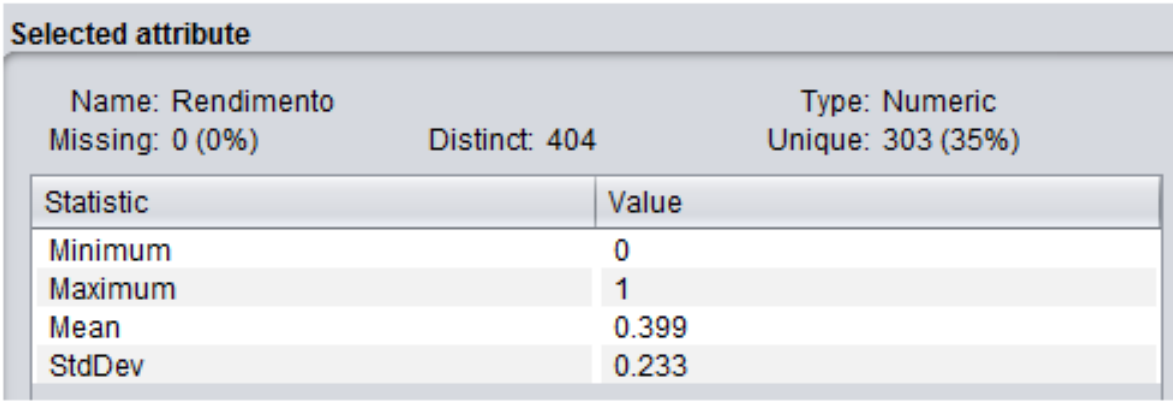


Figura 3.18: Conjunto de dados normalizados

Outro caso verificado que necessita de transformação é a variável data. Esta variável foi transformada juntamente com o arquivo Excel usada para a análise e operações

```
Time taken to build model: 1.31 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9372
Mean absolute error                      0.0524
Root mean squared error                  0.0813
Relative absolute error                 25.8612 %
Root relative squared error             34.8791 %
Total Number of Instances              861
```

Figura 3.19: Resultado do conjunto normalizado

de configurações sobre os dados para um dos formatos de arquivo padrão para as tarefas de mineração de dados, o formato ARFF.

O ARFF aceita basicamente dois tipos de *datatypes* que são String para dados Nominal e *Numeric*. Assim transformamos a variável data para o formato mm/yy/aaaa reconhecido na ferramenta como um tipo data. Esta variável é essencial pois estamos perante dados históricos que se pretende prever para um determinado período, seja mensal ou anual.

3.2.1.4 MINERAÇÃO DE DADOS

A fase de mineração de dados é responsável pela execução dos algoritmos de mineração. Selecionou-se para esta fase dez (10) algoritmos citados no capitulo 2, de predição de funcionalidades diferentes conforme nos estudos dos diferentes autores em problemas relacionados a agricultura apresentados neste estudo. No entanto, aplicaremos estes algoritmos para prever o rendimento das culturas.

Esta fase engloba 4 técnicas e configurações que são necessárias para o sucesso, das quais:

1. Seleção da ferramenta de mineração de dados;
2. Seleção da Técnicas de amostragem para a execução dos algoritmos;
3. Seleção dos algoritmos preditores;
4. Execução dos algoritmos preditores.

3.2.1.4.1 SELEÇÃO DA FERRAMENTA

É importante ter uma ferramenta de mineração de dados para a execução dos algoritmos, pois é nela onde os algoritmos são executados para descobrir os padrões relevantes do processo. Neste estudo a ferramenta WEKA é usada para essa finalidade.

A ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) compõe um ambiente completo para mineração de dados e estão disponíveis diversidades de algoritmos para a tarefa de mineração.

3.2.1.4.2 TÉCNICA DE AMOSTRAGEM

Com a finalização das etapas de seleção, pré-processamento e transformação dos dados antes da aplicação dos algoritmos de mineração de dados nos conjuntos de dados, escolheu-se a técnica de amostragem para avaliar a capacidade de generalização do método a partir dos conjuntos de dados.

A validação cruzada (VC) foi escolhida porque pode ser empregada em problemas onde o objetivo da modelagem é a predição. Ela estima o quão preciso é o método na prática, ou seja, o seu desempenho para um novo conjunto de dados.

Aplicou-se, nos conjuntos de dados, a técnica de VC com número de folhas k=10 que determina o número de iterações a ser executadas sobre o conjunto de dados. Esta técnica divide o conjunto em dez partes sendo que uma parte é usada para teste e os restantes noves são para o treinamento obtendo assim medidas mais confiáveis sobre a capacidade do método.

3.2.1.4.3 SELEÇÃO DOS ALGORITMOS

A seleção dos algoritmos depende do problema em que se está inserido. Uma vez que o objetivo é prever o rendimento das culturas e o conjunto de dados possui valores contínuos, isso significa que se está perante um problema de regressão.

Selecionou-se a partir dos estudos feitos no capítulo 2 da fundamentação teórica, na seção 2.3.2 algoritmos de predição em regressão aplicados no setor agrícola. Dentre os algoritmos estudados apenas dez deles foram considerados para execução nos conjuntos de dados do presente trabalho, dos quais:

1. *Random Forest* (RF): Baseado em conjunto de árvores de decisão (floresta aleatória).

2. *Random Tree* (RT): Árvores aleatórias do tipo estocástico.

3. *Reduced-Error Pruning Tree* (RepTree): Cria várias árvores em diferentes iterações.

4. *M5-Prime* (M5P): Combina uma árvore de decisão convencional com a possibilidade de funções de regressão linear.
5. *K-Nearest Neighbor* (KNN): Baseado no quão similar é um dado do outro.
6. *Bagging* (BG): Melhora a estabilidade e a precisão dos algoritmos de aprendizado de máquina usados na regressão estatística.
7. *Multilayer Perceptron* (MLP): Identifica pequenas seções linearmente separáveis das entradas dos dados. Baseado em redes neurais.
8. *Additive Regression* (AR): Maximiza a qualidade da previsão de uma variável dependente Y a partir de várias distribuições, estimando funções não específicas das variáveis preditoras que são "conectadas" à variável dependente por meio de uma função de link.
9. *Linear Regression* (LR): Estima a condicional de uma variável y, dados os valores de outras variáveis x. Em geral, tem como objetivo tratar de um valor que não se consegue estimar inicialmente.
10. *Regression By Discretization* (RBD): Reduz o número de valores que uma variável contínua assume agrupando-os em um número de intervalos ou posições.

Esses algoritmos selecionados foram executados na ferramenta weka, quando aplicados a técnica de validação cruzada com k=10 em cada conjunto de dados de forma independente. Os algoritmos foram executados usando a padronização do WEKA.

3.2.1.4.4 AVALIAÇÃO DOS MODELOS PREDITIVOS

Cada algoritmo selecionado no estudo, apresenta características particulares de domínio para cada necessidade de experimentação executada, o que demostra efetividade nos resultados. De acordo com Faceli *et al.* [FLG+11] a validação de qualquer técnica de aprendizado de máquina envolve a realização de experimentos controlados em que se demonstre a sua efetividade na solução de diferentes problemas. Os autores acrescentam que é recomendável seguir os procedimentos que garantem a corretude, a validade e a reprodutividade dos experimentos realizados e, mais importante, as conclusões obtidas a partir de seus resultados.

Considerando modelos preditivos a discussão concentrou-se em medidas relacionadas ao desempenho obtido das predições realizadas pelas métricas dos erros baseada em regressão, definidas na seção 2.3.3, entre elas:

1. Erro Médio Absoluto (MAE): Consiste em calcular o residual de cada variável, onde valores residuais negativos e positivos não se anulam.

2. Raiz do Erro Médio Quadrático (RMSE): Baseia-se na média das diferenças entre as predições e observações reais ao quadrado.

Analisaremos também o Coeficiente de Correlação (CC ou R) que mede a relação linear entre as previsões do modelo de regressão e os valores reais, e explica que, quanto maior for o CC maior a correlação existente entre as variáveis.

3.2.1.4.5 EXECUÇÃO DOS ALGORITMOS MINERAÇÃO DE DADOS

Os algoritmos de predição baseado em regressão, são executados nesta seção respeitando todo o procedimento da etapa de mineração anteriormente explicado.

Na ferramenta WEKA, selecionamos por cada execução um conjunto de dados onde aplicamos todas as técnicas mencionadas na seção da mineração de dados. A figura 3.20 apresenta a lista dos atributos que cada conjunto de dados possui sendo este mesmo conjunto constituído de 21 atributos e 861 instâncias.

Na mesma figura 3.20, apresentamos a variável alvo do nosso problema. A variável rendimento, o número oito (8) em azul na imagem. Esta é a variável ou atributo com valores do rendimento da cultura, responsável pelo resultado do valor futuro a ser predito.

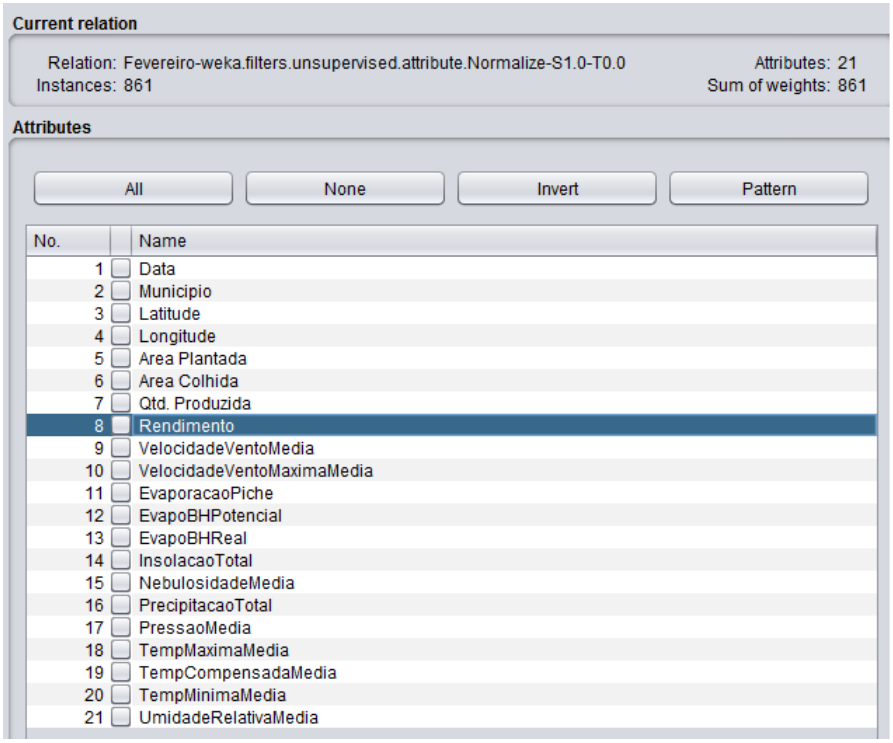


Figura 3.20: Atributos do conjunto de dados

Os conjuntos de dados mensais foram executados obtendo os resultados para cada algoritmo selecionado. Os resultados das execuções são demostrados pelas métricas

que representam o desempenho desses algoritmos. Assim, apresenta-se nas tabelas 3.2, 3.3 e 3.4 como exemplo dos resultados das execuções obtidas dos meses de janeiro a março.

Tabela 3.2: Métricas do mês de janeiro

Algoritmos	**CC**	**MAE**	**RMSE**
trees.REPTree	0,8558	0,0896	0,1217
trees.RandomTree	0,8386	0,0936	0,1301
trees.RandomForest	0,8932	0,079	0,1051
trees.M5P	0,9372	0,0524	0,0813
functions.LinearRegression	0,8605	0,0886	0,12
functions.MultilayerPerceptron	0,0723	0,5062	0,6777
lazy.IBk	0,8482	0,0888	0,1263
meta.AdditiveRegression	0,9079	0,0677	0,0977
meta.Bagging	0,8377	0,0961	0,1275
meta.RegressionByDiscretization	0,9036	0,0694	0,1

Tabela 3.3: Métrica do mês de fevereiro

Algoritmos	**CC**	**MAE**	**RMSE**
trees.REPTree	0,9581	0,0458	0,0774
trees.RandomTree	0,9412	0,0528	0,0921
trees.RandomForest	0,9663	0,0467	0,0705
trees.M5P	0,9556	0,0459	0,0797
functions.LinearRegression	0,955	0,0476	0,0802
functions.MultilayerPerceptron	0,0019	0,3315	0,4194
lazy.IBk	0,9541	0,0462	0,0814
meta.AdditiveRegression	0,9206	0,0703	0,1057
meta.Bagging	0,9475	0,0589	0,0883
meta.RegressionByDiscretization	0,9663	0,0456	0,0693

Tabela 3.4: Métrica do mês de março

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,8807	0,888	0,1183
trees.RandomTree	0,8724	0,0908	0,1224
trees.RandomForest	0,9279	0,0705	0,0927
trees.M5P	0,9394	0,0547	0,0848
functions.LinearRegression	0,8853	0,0875	0,1161
functions.MultilayerPerceptron	-0,0205	0,3245	0,4053
lazy.IBk	0,8714	0,0923	0,1256
meta.AdditiveRegression	0,9049	0,074	0,1053
meta.Bagging	0,8693	0,0943	0,1223
meta.RegressionByDiscretization	0,9187	0,0664	0,0978

3.2.1.5 AVALIAÇÃO/INTERPRETAÇÃO

Não o bastante, a execução dos algoritmos, os resultados obtidos devem ser avaliados. Esta seção é responsável avaliar e interpretar os métodos preditivos gerados dos algoritmos executados.

Esta seção limita-se a avaliar os resultados obtidos das métricas dos erros do MAE e o RMSE. Não interpretaremos os métodos que indicam os valores preditos do rendimento das culturas apesar de este ser o objetivo do processo de DCBD. Limita-se a seguir o objetivo principal do estudo, criar um método de reutilização do processo de DCBD aplicado ao setor agrícola.

Sendo assim, pelos resultados obtidos da execução dos algoritmos apresentados nas seção 3.2.1.4.5, comparou-se cada resultado das métricas dos conjunto de dados dos algoritmos preditores para obter o algoritmo com melhor desempenho para cada mês correspondente. Com os resultados obtidos, podemos ver em azul em cada tabela dos exemplos apresentados para o mês de janeiro e fevereiro, figura 3.2 e 3.3, que o algoritmo **M5P** teve o melhor desempenho e, o mês de fevereiro, 3.4, o algorítimo ***regression by discretization*** apresentou-se com melhor desempenho.

Para além da comparação mensal dos algoritmos, avaliamos a frequência do algoritmo em todos os conjuntos de dados. Essa avaliação feita foi para encontrar o melhor algoritmo do processo que pode ser usado em problemas de predição baseados em regressão no setor agrícola ao pretender prever o rendimento das culturas.

Do resultado das métricas que a tabela 3.6 apresenta, podemos analisar também, graficamente na figura 3.21 para MAE e na figura 3.22 para RMSE, que o algoritmo **M5P** teve melhor desempenho em maior parte dos conjuntos de dados. No entanto, este foi o escolhido como o melhor algorítimo regressor em nosso estudo.

Tabela 3.5: Tabela resumo dos resultados

Conjunto de dados	**Algoritmo**	**CC**	**MAE**	**RMSE**
Janeiro	M5P	0,9372	0,0524	0,0813
Fevereiro	Regression By Discretization	0,9663	0,0456	0,0693
Março	M5P	0,9394	0,0547	0,0848
Abril	Random Forest	0,9567	0,0565	0,0763
Maio	M5P	0,9439	0,0571	0,0837
Junho	M5P	0,9379	0,0551	0,0828
Julho	M5P	0,9477	0,051	0,0813
Agosto	M5P	0,9417	0,059	0,0846
Setembro	M5P	0,9346	0,0576	0,0841
Outubro	M5P	0,9431	0,0511	0,0821
Novembro	M5P	0,9411	0,0497	0,0801
Dezembro	M5P	0,9475	0,0512	0,0848

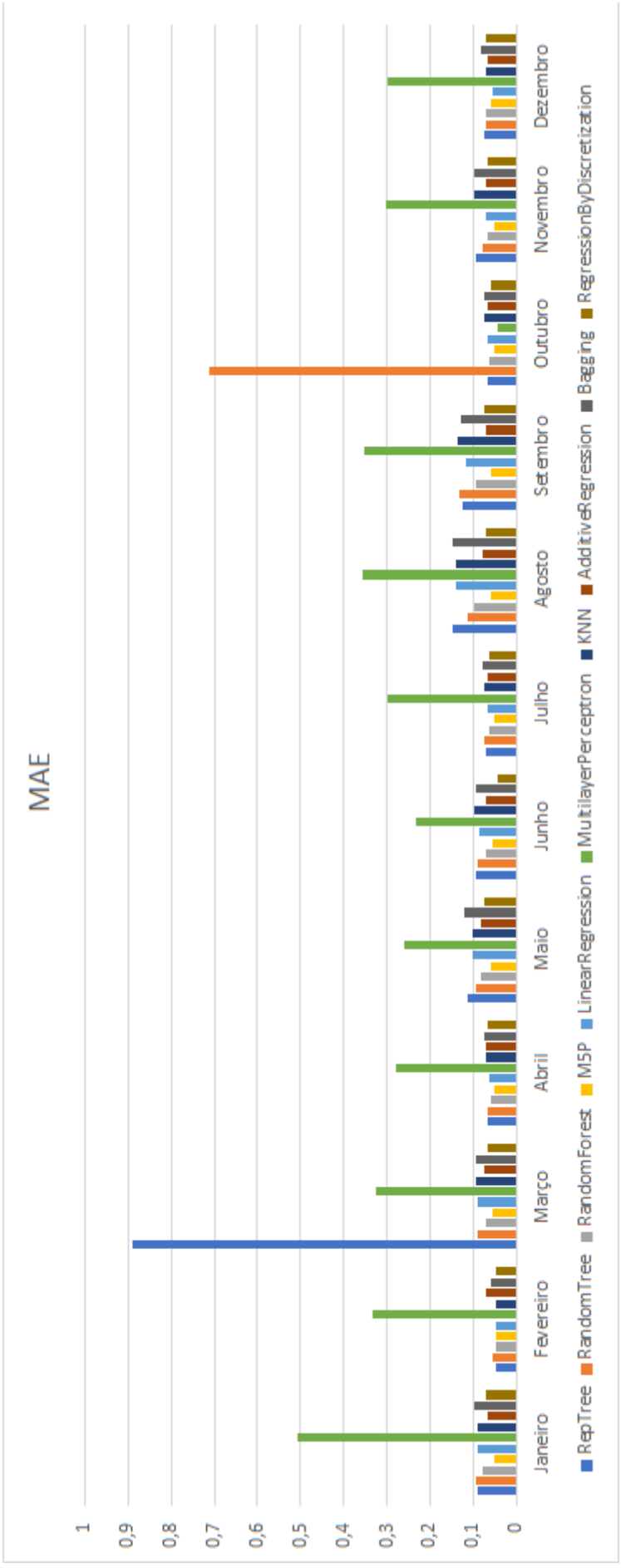


Figura 3.21: Resultados da métrica do MAE dos conjuntos dos dados

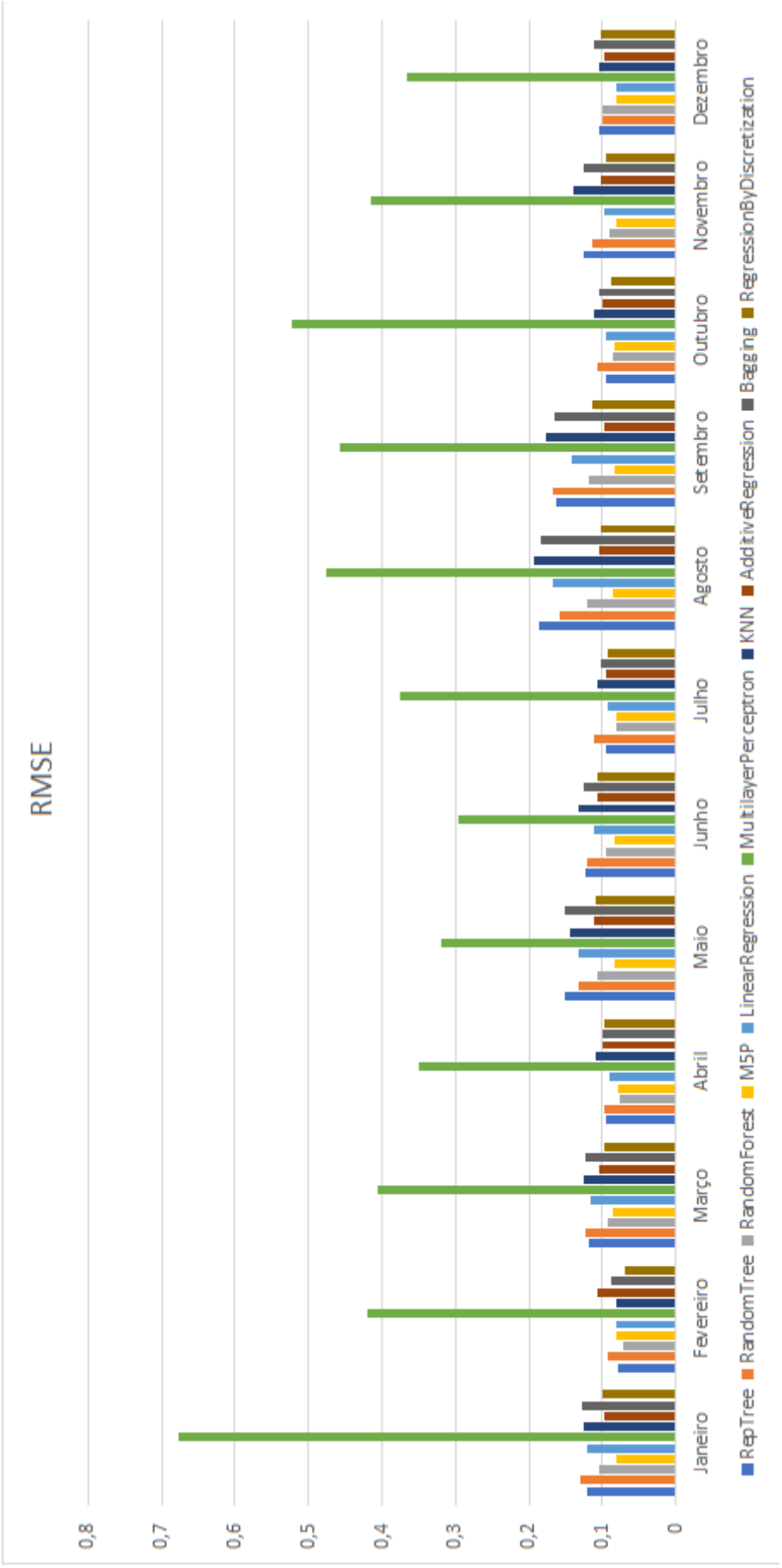


Figura 3.22: Resultados da métrica do RMSE dos conjuntos dos dados

3.2.2 Considerações do Processo de DCBD

A seção 3.2.1 descreveu a execução do processo prévio de DCBD pelo uso de dados da cultura do arroz, executado nos dados colhidos na região do RS no período de 2013 a 2015. Este processo de DCBD, definido pelos autores Fayyad, Piatetsky-Shapiro e Smyth, teve como objetivo principal identificação das tarefas realizadas sobre o processo que podem ser reutilizadas em novos casos.

Identificou-se também, durante a execução do processo, os algoritmos de predição baseada em regressão para predizer cada mês e para predizer anualmente o rendimento das culturas. Assim, tratando-se de um processo de DCBD conclui-se na base da execução do processo, prévio que o algoritmo M5P é o que aparece com maior frequência, o que torna o escolhido para resolver problemas de regressão no conjunto de dados do setor agrícola.

Na seção pode-se verificar que durante a execução do processo, procurou-se analisar os dados desde a primeira etapa do processo por forma a identificar as tarefas a serem realizadas e a parametrização ocorrida em cada fase. Porém, o processo teve muito detalhe para identificação das configurações realizadas, na obtenção do bom resultado do método onde não apenas foram consideradas tarefas da metodologia proposta por Fayyad, Piatetsky-Shapiro e Smyth, mas também tarefas que se mostraram necessárias durante a modelagem dos dados. Estas tarefas executadas, tornam um diferencial aplicado no processo de DCBD e essenciais para a etapa do método pois é através dele que essas tarefas são identificadas para o reúso em outros processo se tornando assim, importante a sua execução.

3.3 BASE DO CONHECIMENTO DO PROCESSO

Esta seção descreve as tarefas que norteiam o desenvolvimento do método para o apoio à reutilização do processo de DCBD aplicado no setor agrícola. Como forma apoiar a execução do processo de DCBD, a base de conhecimento é composta pelas fases que o norteiam, composta pelas tarefas aproveitadas do processo prévio executado, apresentadas na figura 3.23.

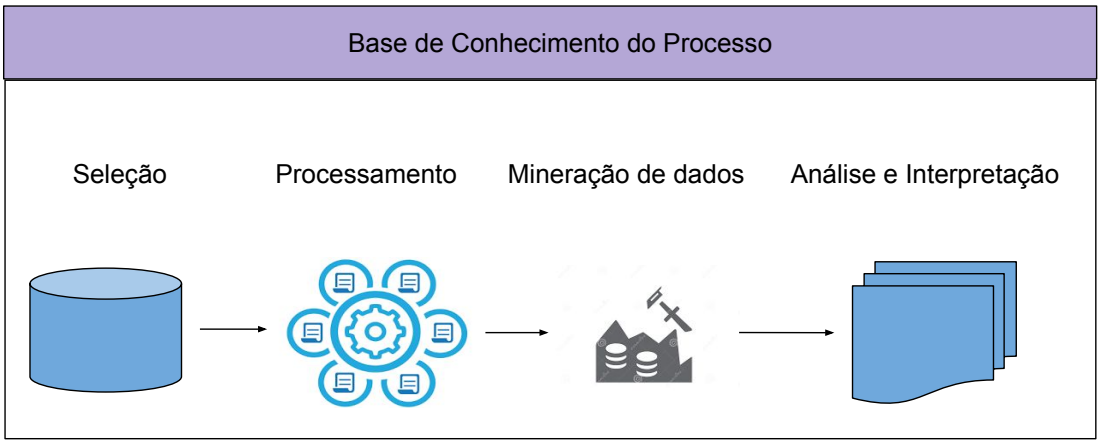


Figura 3.23: Base de Conhecimento do Processo

Esta base de conhecimento é um repositório de tarefas que poderá diminuir proporcionalmente o tempo de execução de um processo de DCBD que apresentaremos no capitulo 4 do teste da solução. Nela estão armazenadas todas as experiências adquiridas nos processos de DCBD executados sobre o processo prévio. Em suma, a Base de Conhecimento possui os seguintes benefícios para um processo de DCBD:

- Agilidade na solução para a execução do processo de DCBD;
- O conhecimento permanece no setor para a disposição de todos. O uso não será apenas para o analista do domínio;
- Facilidade de uso pelos novos usuários devido a descrição detalhada do conteúdo homologado.

Dentre as fases que compõem a figura 3.23 da base de conhecimento, descrevemos nas suas fases a seguir, os requisitos para a reutilização:

1. Fase da Seleção

   Esta fase identifica as informações que devem efetivamente ser consideradas, para o processo de DCBD, aplicadas no setor agrícola. A fase auxilia o usuário de forma

mais simples possível na obtenção de informações necessárias sem recorrer a conhecimentos complexos, simplificando deste modo grande parte do conhecimento exigido.

Uma facilidade oferecida é de que permite uma visão geral das instâncias e dos atributos que o conjunto de dados deve possuir. O número de instância não é constante pois depende do número de municípios que produz a cultura a ser processada.

2. Fase do Processamento

   Esta fase propõe as tarefas do processo para apoiar a sua execução evitando retrabalho de detalhes das configurações realizadas no processo prévio. Essas configurações são reusadas em casos semelhantes ao processo a ser desenvolvido para que o usuário em mineração de dados possa tomar as melhores decisões na execução do processo.

   Em suma, esta etapa auxilia o usuário nas tarefas de pré-processamento e transformação de dados selecionados, avaliando individualmente cada coluna da tabela e sugerindo o que deve ser feito para deixar a coluna apta para a mineração através da parametrização das tarefas.

3. Fase da Mineração de dados

   A mineração de dados é a principal etapa do processo de DCBD que ocorre na busca efetiva por conhecimentos. Para alcançar o objetivo é necessário que se aplique algoritmos para prever o rendimento das culturas.

   A proposta desta etapa é auxiliar o usuário a dispensar a necessidade de conhecimento detalhado sobre os algoritmos disponíveis e de detalhes sobre seus parâmetros de configuração, limitado-se apenas na sua aplicação no processo.

4. Fase da Avaliação e Interpretação

   Esta etapa auxilia o usuário na execução das atividades que compreendem a mineração de dados, avaliando as métricas de predição e interpretando os modelos obtidos dos algoritmos de executados.

3.3.1 REQUISITOS PARA APOIAR O PROCESSO DE DCBD NO SETOR AGRÍCOLA

A classificação detalhada nas seções subsequentes, resume-se na descrição das contribuições e experiências consideradas durante a execução do processo prévio. Essas contribuições e experiências são consideradas como configurações catalogadas na base de conhecimento que é o método guiado para a reutilização do processo de DCBD para o setor agrícola.

3.3.1.1 FASE I: SELEÇÃO

Atividade da fase: Busca pelos dados.

Descrição da atividade: Deve-se ter os dados agrícolas, meteorológicos/climáticos, coordenadas geográficas relacionados aos municípios da região em pesquisa.

Estratégia da fase: Seleção dos atributos listados:

- Agrícola: área plantada, área cultivada, rendimento da cultura e a quantidade produzida.
- Meteorológicos: Todos os possíveis dados climáticos sendo que os principais: Temperaturas (minímas, média e máximas), umidade relativa do ar, precipitação, insolação e velocidade do vento;
- Coordenadas Geográficas: Dados das coordenadas geográficas da região em estudo, em graus decimais.

3.3.1.2 FASE II: PROCESSAMENTO

Tabela 3.6: Processamento

Atividade	**Descrição**	**Especificação**
Ajustar os dados agrícolas.	Os dados agrícolas devem estar em meses. Usando a regra de 3 simples determinar o valor mensal de cada município da região.	Aplicar a fórmula: $Vmês = \frac{total\_anual \times percentagem\_mensal}{100}$
Integrar os conjuntos de dados.	Integrar os conjuntos de dados selecionados em um único conjunto de dados.	Usando os municípios como a variável correspondente da unificação dos conjuntos de dados.
Imputar valores dos municípios sem estações meteorológicas	Calcular a distância latitudinal usando dados das coordenadas dos municípios para encontrar os municípios vizinhos dos municípios com estações meteorológicas e calcular o valor a ser imputado.	Fórmula da distância latitudinal: $DL = \|LA_f - LA_i\|$; Fórmula do valor a imputar: $Valor\_inter = \frac{soma\_valores * total\_observacoes}{total\_observacoes}$
Imputar valores faltantes	Devem ser resolvidos a falta de dados em campos sem informação.	Preenchimento usando o caractere "?" em campos vazios.
Estratificar o conjunto de dados.	Conjunto de dados possuem informação mensal desbalanceada.	Estratificar pelos 12 meses ignorando os anos.
Transformar dados do conjunto.	Colocar os dados padronizados	Aplicar o filtro *normalize* do WEKA
Transformar a data e o arquivo	A data deve estar no formato data e no arquivo ARFF.	Uso da ferramenta WEKA para auxilio da transformação do arquivo do conjunto de dados para a extensão .arff.

3.3.1.3 FASE III: MINERAÇÃO DE DADOS

Atividade: Executar os algorítimos de mineração de dados para extrair o conhecimento existente nos dados.

Descrição da atividade: Aplicar sobre os conjuntos de dados o algoritmo M5P ou os dez algoritmos preditores listados:

1. *Random Forest* (RF);

2. *Random Tree*; (RT)
3. *Reduced-Error Pruning Tree* (REPTree);
4. *M5-Prime* (M5P);
5. *K-Nearest Neighbors* (KNN);
6. *Bagging* (BG);
7. *Multilayer Perceptron* (MLP);
8. *Additive Regression* (AR);
9. *Linear Regression* (LR);
10. *Regression By Discretization* (RBD).

Estratégia: Aplicar com ajuda da ferramenta weka a técnica de validação cruzada com k=10 e executar os algoritmos de regressão.

3.3.1.4 FASE IV: ANÁLISE E INTERPRETAÇÃO

Atividade: Analisar os resultados e interpretar os modelos obtidos durante a execução dos algoritmos de mineração de dados.

Descrição: Análise dos resultados das métricas dos erros, dados pelas fórmulas:

- $MAE = \frac{1}{n}\sum_{t=1}^{n}|y_i - \hat{y}_i)|$;
- $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y}_i)^2}$;
- $C = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$

E interpretação dos modelos correspondentes a cada algoritmo.

Estratégia: Na análise, considerar o menor valor dos erros das métricas para avaliar o desempenho do algoritmo. Interpretar o modelo gerado pelo algoritmo.

3.3.2 Considerações da Base de Conhecimento

Esta etapa do método apresentou as tarefas que devem ser executadas no processo de DCBD, permitindo que um usuário comum use-o de forma simples sem se preocupar com detalhes de configuração para sua execução.

A Etapa da base de conhecimento apresenta as tarefas distribuídas em fases que o compõem. Esta distribuição facilita na execução do processo, dando a conhecer as fases do processo em execução o que permite o usuário entender como ela deve ser aplicada e em que fase de execução se encontra. Esta metodologia documentada aplicada nesta etapa não apenas proverá a agilidade no processo como também permite que um usuário comum possa executar e compreender o método com facilidade.

3.4 A REUTILIZAÇÃO DO PROCESSO DE DCBD

No âmbito da execução do processo reutilizado, esta seção executa o processo reutilizado a partir da base de conhecimento criada com a execução do processo inicial sobre a cultura do arroz. Como forma de testar a reutilização, o processo será executado sobre a cultura do feijão no Estado do Rio Grande do Sul.

Procuramos testar em uma cultura com características diferentes da cultura do arroz na mesma região como forma de validar a aplicabilidade da metodologia. Em relação ao clima, segundo Pereira Gris [PGM+14], a temperatura ideal para a planta (feijão) se situa entre 10ºC e 25ºC, embora a cultura da leguminosa também possa ser feita em temperaturas acima de 35ºC, escolhendo-se variedade adequada sob regime de irrigação. Se houver chuvas com uma precipitação pluviométrica de mais ou menos 100mm na época do plantio e do crescimento do feijão, tanto melhor para a cultura.

De acordo com a base de conhecimento criada, a execução do novo processo evitará configurações iniciais que dispendem muito tempo durante sua execução e testará a sua execução em outras culturas do mesmo domínio.

3.4.1 SELEÇÃO

Uma vez que a implementação será no estado do RS, onde foi realizada a execução do processo inicial, alguns dados colhidos foram reaproveitados, como o caso da lista das coordenadas geográficas e da lista das estações meteorológicas. Também, reaproveitou-se fontes de busca para a seleção dos dados de produção da cultura do feijão. Uma vez que esses dados são provenientes das mesmas fontes, eles apresentam o mesmo tipo de configuração encontrado no processo inicial.

Com base nas tarefas catalogadas na base de conhecimento, tornou este processo mais flexível em relação ao primeiro processo executado uma vez que ele nos fornece informações do tipo de dados que é necessário para o processo. Pode acontecer, no entanto, que os dados necessários para algumas regiões ou algumas culturas não se encontrem nas fontes de dados indicadas no estudo, para o caso, importa saber que tipo de dados são necessários facilitando deste modo a busca, tornando fácil e ágil.

Os dados selecionados da cultura do arroz apresentam a mesma descrição dos dados da cultura do arroz apresentados no capitulo 3, tabela 3.1. O diferencial encontrado nos dados selecionados foi a quantidade de instâncias em que a cultura de arroz o conjunto de dados inicial continha 21 atributos e 12672 instâncias. A cultura do feijão possui mais instâncias em relação ao arroz, de 22 atributos e 22044 instâncias. Este diferencial

em quantidade de dados não interfere no processo e pode ser uma situação comum para diferentes culturas.

3.4.2 PROCESSAMENTO

Depois da obtenção dos dados, a fase seguinte é o processamento, como indica a base de conhecimento. Esta é a fase mais demorada em todos processos de DCBD, incluindo o processo reutilizável. Embora as configurações estejam catalogadas ou anotadas para o reúso, pode ser que a fase passa a ser mais rápida, o que pode agilizar bastante o tempo de processamento.

Como forma de mostrar a agilidade que esta fase pode beneficiar-se, a partir do conhecimento das configurações necessárias obtidas da execução do processo prévio, foram reutilizadas as configurações anotadas na base de conhecimento. A tabela 3.7 apresenta as configurações realizadas nesta fase guiada pela base de conhecimento.

Tabela 3.7: Processamento

Atividade	**Descrição**	**Ferramenta**
Análise dos atributos	Analisados cada atributo contido em cada conjunto de dados.	Microsoft Excel
Integração dos conjuntos de dados	Correspondência entre os municípios produtores tendo em conta o período selecionado.	Microsoft Excel
Interpolação dos dados	Achar as estações mais próximas entre os municípios e calcular a média para preencher os campos vizinhos. Uso das coordenadas geográficas para achar a distância entre os municípios.	Microsoft Excel
Tratamento de dados em falta	Imputação pelo caractere "?" para casos desconhecidos.	Microsoft Excel
Análise de correlação	Entender graficamente como as variáveis estão correlacionadas.	Excel ou RATTLER[1]
Transformação do atributo data	os atributos da data encontram-se no formato nominal converte-se para o formato data.	WEKA
Transformação do conjunto de dados	converter o arquivo .csv em .artff.	WEKA
Transformação do conjunto de dados	Normalização pela faixa de variação 0 a 1.	WEKA

[1]Uma interface gráfica do usuário para mineração de dados, usando R.

Este processo foi executado, respeitando toda a configuração da base de conhecimento, pois ela definirá o resultado da execução da fase da mineração.

O tempo de execução deste processo é relativo a cada situação ou de acordo com o problema a ser resolvido. Esta permite uma fácil reutilização das configurações (ajustes, fórmulas e adaptações) a serem executadas que tornou o processo significativamente ágil em relação ao processo prévio, devido ao conhecimento que ficou disponível na base de conhecimento para o reuso.

Embora não incluímos a tarefa de análise das variáveis do conjunto nas atividades da base de conhecimento, realizamos, como forma de teste, para comprovar que essas variáveis selecionadas são importantes para o problema. Durante a análise da correlação das variáveis do conjunto, verificou-se que a cultura de feijão apresenta correlação entre as variáveis de forma independente com as variáveis meteorológicas sendo elas positivas ou negativas. Este resultado explica que as culturas possuem características diferentes em relação ao clima ou regiões de plantação mas que apesar disso elas apresentam uma correlação.

Essa diferença nas características das variáveis e no comportamento, pode de certa forma, influenciar no comportamento dos algoritmos de mineração de dados que podemos ver mais adiante. A figura 3.24 apresenta o resultado das correlações entre as variáveis da cultura de feijão.

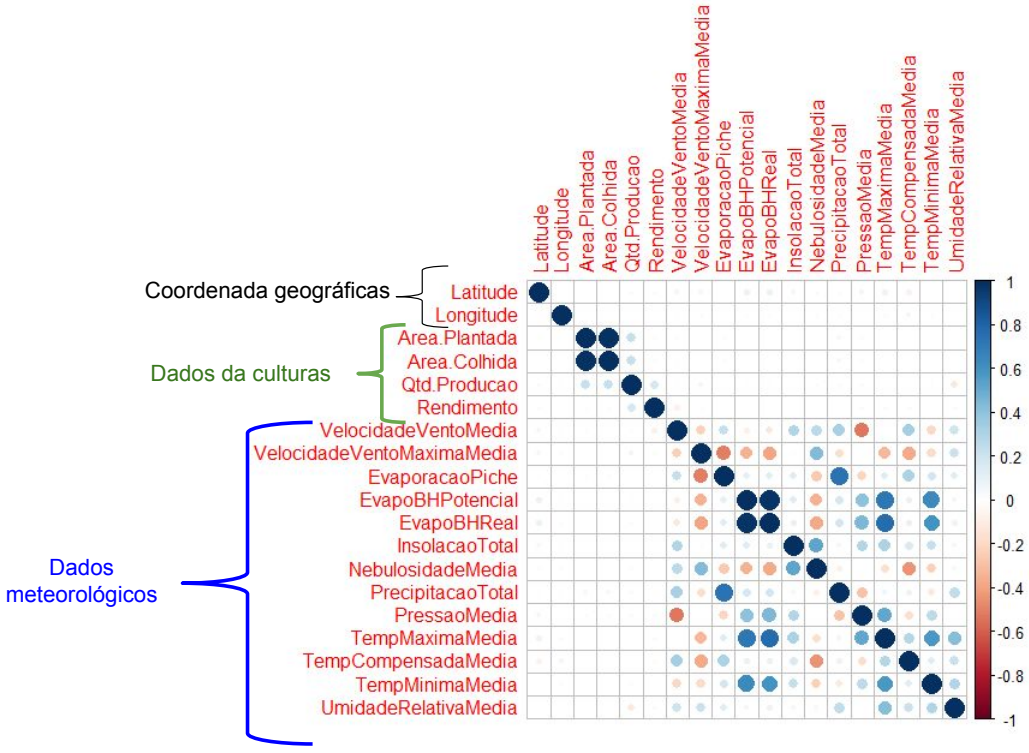


Figura 3.24: Correlação entre as variáveis da cultura do feijão

O mesmo comportamento encontra-se nos conjuntos de dados estratificados por mês. Como podemos ver nas figuras 3.25 e 3.26 para os meses de janeiro a abril e as outras, de maio a dezembro no apêndice E em que se apresenta o mesmo comportamento. As variáveis desses conjuntos se correlacionam de forma independente o que explica que apesar da existência ou não de altas ou baixas temperaturas, esta cultura é resistente a variação climática. Uma vez que o clima tem comportamento diferente para cada mês em cada ano, esta verificação poderá ajudar a medir as influências do clima na cultura de feijão.

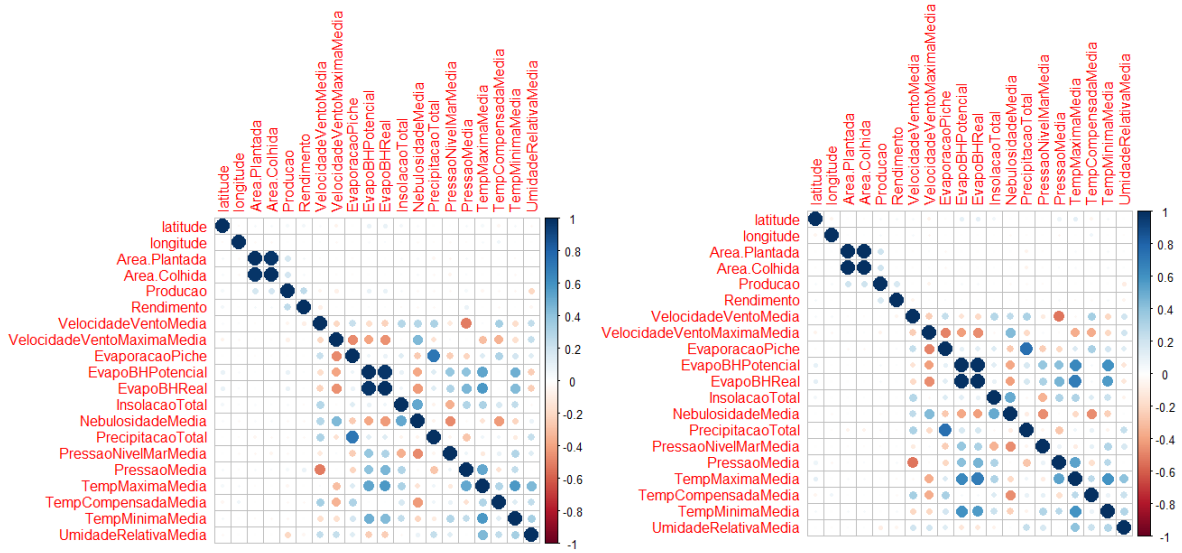


Figura 3.25: Correlação do mês de janeiro e fevereiro do feijão

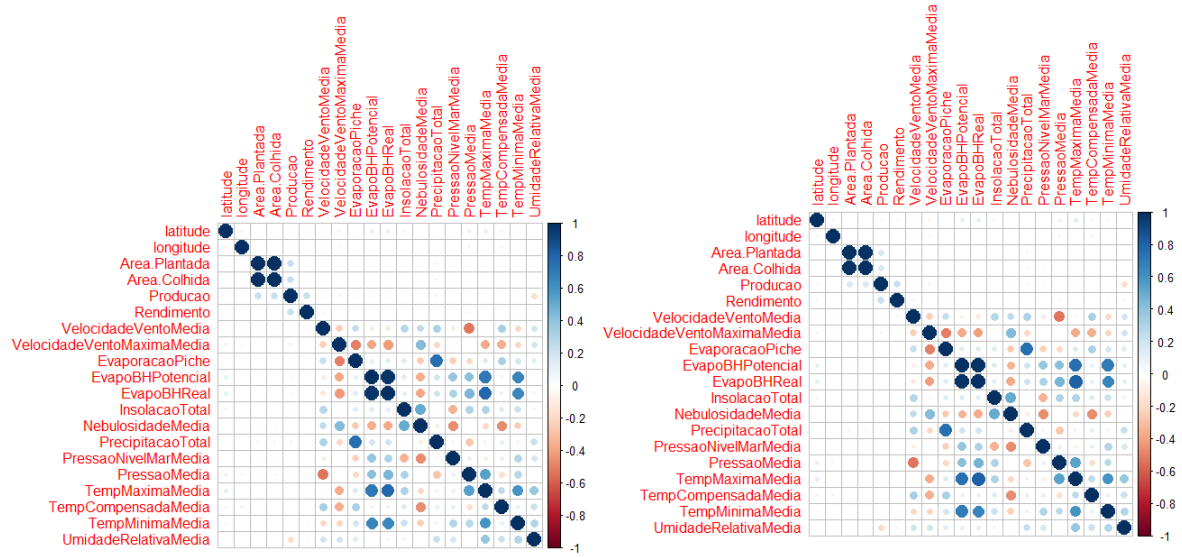


Figura 3.26: Correlação do mês de março e abril do feijão

3.4.3 MINERAÇÃO DE DADOS

Nesta fase foram executados os 10 algoritmos de previsão baseados em regressão, selecionados, para encontrar os algoritmos com melhor desempenho, conforme explicados nos capítulos anteriores.

Como forma de analisar se os algoritmos selecionados para o problema de regressão, para prever o rendimento das culturas, tem um bom desempenho ou apresenta-se com bom desempenho em casos reutilizáveis, executamos sobre a cultura do feijão todos algoritmos ao invés de apenas executar o algoritmo indicado, o M5P.

Com esta execução, comparou-se os resultados das métricas dos erros obtidos por forma a obter um algoritmo padrão, validado para o modelo reutilizável do processo de DCBD. A seguir apresenta-se os exemplos dos resultados das execuções obtidas na tabela 3.8 para o mês de janeiro, na tabela 3.9 para o mês fevereiro e na tabela 3.10 para o mês de março.

Tabela 3.8: Métrica de janeiro para a cultura do feijão

Algoritmos	**CC**	**MAE**	**RMSE**
trees.REPTree	0,5334	0,0983	0,1285
trees.RandomTree	0,4789	0,1093	0,1453
trees.RandomForest	0,6694	0,0825	0,1096
trees.M5P	0,5355	0,0766	0,163
functions.LinearRegression	0,5841	0,0942	0,1243
functions.MultilayerPerceptron	0,0017	0,1453	0,1798
lazy.IBk	0,5089	0,1097	0,1431
meta.AdditiveRegression	0,5119	0,097	0,1256
meta.Bagging	0,567	0,095	0,124
meta.RegressionByDiscretization	0,5668	0,0953	0,1251

Tabela 3.9: Métrica de fevereiro para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,713	0,0775	0,1098
trees.RandomTree	0,554	0,0951	0,1356
trees.RandomForest	0,7526	0,074	0,102
trees.M5P	0,7269	0,074	0,1073
functions.LinearRegression	0,7179	0,0759	0,1091
functions.MultilayerPerceptron	0,0353	0,1332	0,1656
lazy.IBk	0,6421	0,0881	0,129
meta.AdditiveRegression	0,3963	0,1095	0,1405
meta.Bagging	0,698	0,0783	0,1106
meta.RegressionByDiscretization	0,7105	0,0783	0,1096

Tabela 3.10: Métrica de março para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,5975	0,0934	0,1242
trees.RandomTree	0,6234	0,0901	0,1249
trees.RandomForest	0,7585	0,0726	0,099
trees.M5P	0,7324	0,0706	0,1029
functions.LinearRegression	0,7369	0,0713	0,1028
functions.MultilayerPerceptron	0,0154	0,1388	0,1698
lazy.IBk	0,6174	0,0965	0,1308
meta.AdditiveRegression	0,497	0,0991	0,129
meta.Bagging	0,5916	0,0922	0,1226
meta.RegressionByDiscretization	0,5943	0,0936	0,1239

3.4.4 ANÁLISE DOS RESULTADOS

Ao analisar os resultados obtidos das execuções dos algoritmos foram encontrados casos em que um determinado mês as métricas diferenciaram-se em relação aos algoritmos. Por exemplo, no mês de janeiro representado pela tabela 3.8, o algoritmo M5P foi melhor para o MAE enquanto que o algoritmo Random Forest foi melhor para a métrica

RMSE representado a verde. O mesmo comportamento deu-se para o mês de março apresentado pela cor verde na tabela 3.10. Para o mês de fevereiro, tabela 3.9, o algoritmo **Random Forest** mostrou-se com melhor desempenho nas duas métricas em uso, assinalado pela cor verde.

No caso em que os resultados das métricas foi de dois algoritmos, observou-se a velocidade da execução destes algoritmos nos respetivos meses para a escolha do melhor desempenho para representar o mês. Nos exemplos, o mês de março o algoritmo M5P levou 10,81 segundos para construir o modelo enquanto que o *Random Forest* levou 2,05 segundos na sua construção conforme apresentado nas figuras 3.27 e na figura 3.28. Assim, o algoritmo Random Forest tornou-se o melhor para este mês.

```
Time taken to build model: 10.81 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5355
Mean absolute error                      0.0766
Root mean squared error                  0.163
Relative absolute error                 67.3959 %
Root relative squared error            111.4027 %
Total Number of Instances             1837
```

Figura 3.27: Resultado das métricas do M5P.

```
Time taken to build model: 2.05 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6694
Mean absolute error                      0.0825
Root mean squared error                  0.1096
Relative absolute error                 72.643  %
Root relative squared error             74.953  %
Total Number of Instances             1837
```

Figura 3.28: Resultado das métricas do Random Forest.

Após as análises mensais em cada conjunto de dados, gerou-se a tabela 3.11 contendo os resultados mensais dos algoritmos que se mostram com melhor desempenho em cada mês. A partir dos resultados obtidos, analisamos a frequência dos algoritmos em todos os conjuntos por forma a validar o modelo para o problema de regressão usado na previsão do rendimento das culturas. Do resultado obtido podemos concluir que os dois

algoritmos *Random Forest* e M5P podem ser usados para prever o rendimento das culturas embora, quando comparados o tempo de execução, o algoritmos Random Forest mostrou mais eficiência. Esse resultado pode também ser analisado graficamente em relação aos erros de acordo com a figura 3.29 para a métrica MAE e figura 3.30 para a métrica RMSE.

Tabela 3.11: Tabela resumo dos resultados

Conjunto de dados	**Algoritmo**	**CC**	**MAE**	**RMSE**
Janeiro	RandomForest	0,0694	0,0825	0,1096
Fevereiro	RandomForest	0,7526	0,074	0,102
Março	RandomForest[1]	0,7585	0,0726	0,099
Abril	RandomForest	0,7368	0,0754	0,1013
Maio	M5P	0,7482	0,0703	0,0971
Junho	RandomForest	0,7504	0,0709	0,0961
Julho	RandomForest[1]	0,7295	0,0765	0,1052
Agosto	M5P	0,7827	0,0701	0,0994
Setembro	M5P	0,7413	0,0752	0,1012
Outubro	M5P	0,694	0,0752	0,1029
Novembro	RandomForest	0,7544	0,0718	0,0979
Dezembro	M5P[1]	0,722	0,0701	0,1015

[1]Algoritmo selecionado de acordo com a eficiência no tempo de execução do modelo.

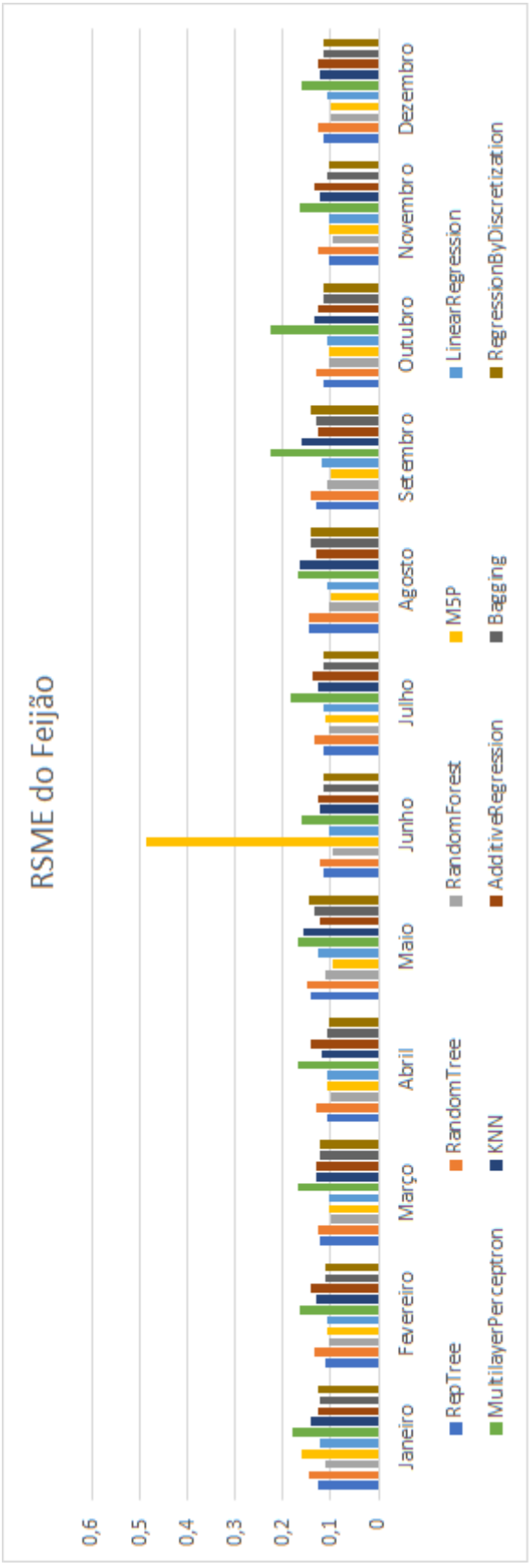


Figura 3.29: Resultados das métricas do RMSE do Feijão

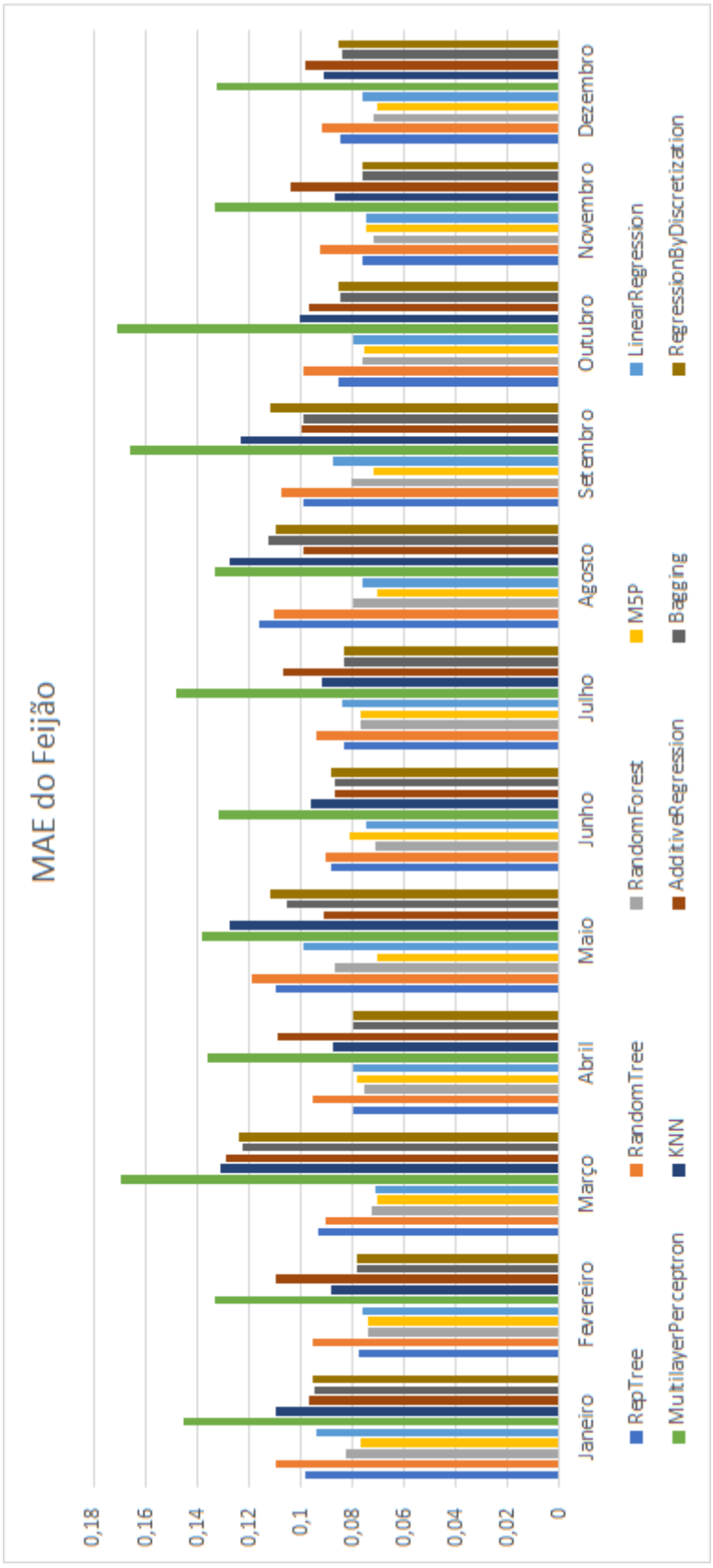


Figura 3.30: Resultados das métricas do MAE do feijão

3.4.5 EXECUÇÃO DO MODELO M5P

Ainda no âmbito da execução do processo reutilizado, esta seção executa o algoritmo M5P sobre os conjuntos dos dados. O algoritmo de predição baseado em regressão, M5P, é unicamente executado sobre o conjunto de dados da cultura do feijão como sendo o algoritmo escolhido para resolver problemas de predição do rendimento das culturas.

Durante a execução do processo de DCBD prévio executado com dados da cultura do arroz, o algoritmo M5P foi o escolhido em relação aos outros pelo melhor desempenho em relação aos outros algoritmos que culminou em mais frequência determinando assim a sua escolha. Executou então, este algoritmo nos doze conjuntos de dados onde obteve-se os resultados da tabela 3.12 para as métricas dos erros do MAE e do RMSE.

Tabela 3.12: Resultado do M5P

CONJUNTO DE DADOS	**CC**	**MAE**	**RMSE**
Janeiro	0,5355	0,0766	0,163
Fevereiro	0,7269	0,074	0,1073
Março	0,7324	0,0706	0,099
Abril	0,7044	0,0785	0,1092
Maio	0,7482	0,0703	0,0971
Junho	0,7175	0,0814	0,4881
Julho	0,714	0,077	0,1052
Agosto	0,7827	0,0701	0,0994
Setembro	0,7413	0,0716	0,1012
Outubro	0,694	0,752	0,1029
Novembro	0,7227	0,0748	0,1046
Dezembro	0,722	0,0701	0,1015

A partir dos resultados obtidos nos conjuntos de dados pode-se verificar bons resultados das métricas o que comprova que o algoritmos M5P pode ser utilizado para pre-

dizer o rendimento das culturas. Também, pode-se verificar pelo modelo de árvore da figura 3.31 obtido a partir do algoritmo M5P, que os atributos selecionados como a data, as coordenadas geográficas, dados da cultura e meteorológicos são importantes e considerados para prever o rendimento das culturas.

Pode-se observar também, por exemplo, que na árvore a variável das coordenadas, latitude, é a raiz e cada nó interno corresponde as outras variáveis contidas no conjunto de dados. Em cada nó pode-se observar o valor do rendimento predito considerado promissor o que pode ajudar o setor agrícola a partir dele a encontrar a melhor situação a considerar em suas atividades.

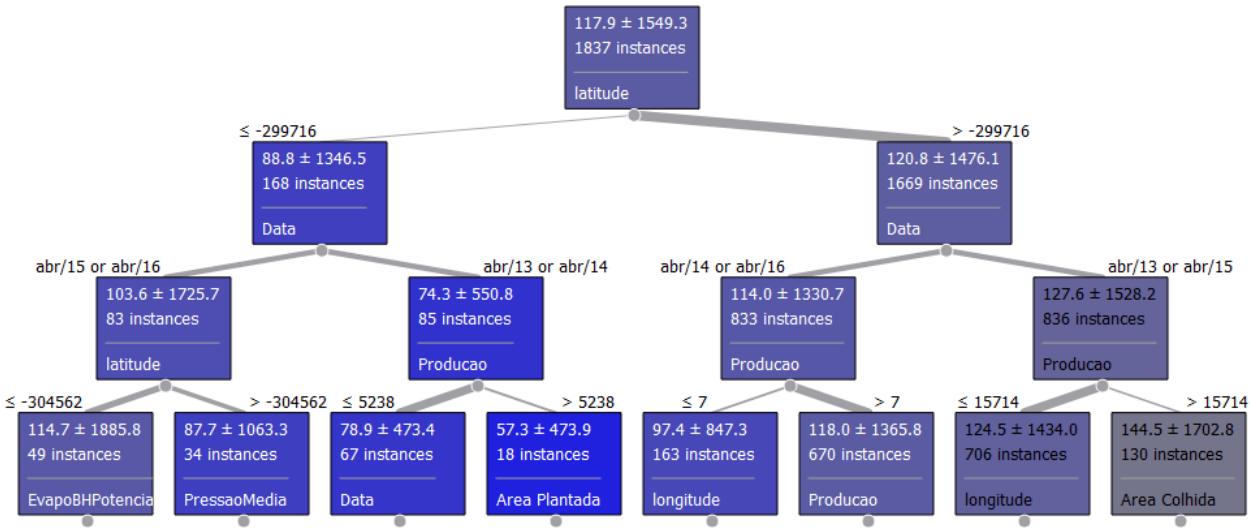


Figura 3.31: Árvore de regressão

3.4.6 CONSIDERAÇÕES DA REUTILIZAÇÃO

Durante a execução da reutilização do processo de DCBD realizado sobre a cultura do feijão, com ajuda da base de conhecimento, verifico-se que houve uma significância agilidade na execução do processo.

Essa agilidade deveu-se à disponibilidade das configurações catalogadas na base de conhecimento que facilitaram a execução do novo processo. Foi possível notar, durante a execução do processo reutilizável, que um usuário qualquer pode executar o processo de DCBD usando como guia a base de conhecimento, desde que este esteja disponível para o seu acesso.

Com os resultados obtidos durante esta execução do processo reutilizado, foi possível verificar o comportamento diferenciado dos algoritmos. Esta situação pode suceder

em outros casos devido a características diferentes das culturas, relacionando também a diferença climática que pode influenciar nas culturas. No entanto, foi possível encontrar durante a execução dos algoritmos, na cultura do feijão, o algoritmos de regressão M5P igual ao encontrado no arroz.

Embora, o algoritmo *Random Forest* tenha sido escolhido na cultura do feijão com melhor desempenho, o algoritmo M5P mostrou-se eficiente na sua execução em todos os conjuntos de dados e com este desempenho apresentado, pode-se concluir que ambos podem ser considerados na previsão do rendimento das culturas.

Do modelo gerado a partir do algoritmo M5P observou-se que as variáveis selecionadas para o problema de predição do rendimento das culturas são consideradas importantes pois geram resultados de rendimento promissores para o setor.

3.5 CONCLUSÃO DO MÉTODO

Executado o método de reutilização de DCBD, foi provado através de teste da sua execução em etapas que compõem o método. Numa primeira execução do teste do método este se demostrou reutilizável.

A primeira etapa do método, processo de DCBD, foi executado usando dados da cultura do arroz. A segunda etapa, Base de Conhecimento, catalogou as configurações executadas na etapa do processo de DCBD onde os dados foram modelados de acordo com as necessidades que estes apresentavam para execução do processo. Esta etapa permitiu uma descrição das fases que o compõem por forma que um suário comum possa interpretar e executar um processo com facilidade.

A última etapa, a reutilização, executa as tarefas do processo de DCBD usando a base de conhecimento para apoiar a sua execução. Desta forma, esta etapa torna-se de fácil uso uma vez já conhecidas as configurações que as tarefas devem modelar nos dados sem se preocupar com detalhes maiores na sua execução. Assim também, permite agilidade na execução devido a esta disponibilidade que a base de conhecimento disponibiliza para a execução do processo.

Através da execução do método, foi possível verificar que mesmo tendo dados provenientes de outras regiões ou dados de outras culturas, do domínio temporário, eles poderão sofrer algumas modificações que não foram identificadas neste processo. Isto entende-se que dados das culturas de outras regiões podem apresentar configurações diferentes o que no entanto, o método poderá ser executado do início e somente reutilizar em caso similares. Pode-se também, para além dos dez algoritmos sugeridos, aplicar apenas o algoritmo M5P em problemas de regressão para prever o rendimento das culturas.

4. TESTE DA SOLUÇÃO

Para testar a metodologia proposta, foram feitas duas análises: uma análise qualitativa do que foi feito em cada um dos casos e uma análise quantitativa do tempo dispendido em cada processo. Essas análises foram comparadas em termos da sua eficácia que garante a vantagem do teste da solução.

O tempo de execução da operação é uma das maiores vantagens da solução, uma vez que tem como objetivo otimizar o conjunto de tarefas que são executadas de forma independente uma da outra. As análises qualitativas vão complementar as análises quantitativas a partir da descrição detalhada do que foi configurado e executado em cada tarefa do processo, contribuindo assim, para agilização do tempo no processo reutilizado.

Na tabela 4.1 estão listadas as principais tarefas realizadas sobre a cultura do arroz que foi o processo prévio responsável pela determinação das configurações a serem aprendidas para uso em novos processos.

Tabela 4.1: Principais tarefas realizadas durante a execução do processo inicial

Nº de tarefas	**Tarefas realizadas**
01	Seleção dos dados
02	Distribuição mensal por município no conjunto de dados agrícolas.
03	Integração das bases de dados
04	Interpolação das estações meteorológicas.
05	Tratamento dos dados faltantes.
06	Correlação das variáveis da base de dados.
07	Estratificação da base de dados.
08	Normalização dos atributos.
09	Execução dos algoritmos preditores.
10	Avaliação dos resultados.

Para aferir a eficiência da solução testada, de acordo com as tarefas realizadas sobre a cultura do arroz, foi utilizado um conjunto de dados da cultura de feijão, em que as referidas culturas foram submetidas ao processo de DCBD. Este novo processo foi executado com base no uso da base de conhecimento aprendido, método desenvolvido para a reutilização do processo de DCBD.

4.1 ANÁLISE QUALITATIVA DO PROCESSO

Durante a execução do processo de DCBD, as tarefas foram configuradas como forma de adequá-las para a extração do conhecimento. As configurações feitas são usadas para compor a base de conhecimento e reutilizadas em um novo processo de DCBD no setor agrícola, sobre culturas temporárias.

Esta seção faz uma análise qualitativa das configurações e tarefas exercidas no processo prévio e o diferencial aplicado no processo reutilizado. A ideia base desta análise é mostrar, de forma detalhada, o tratamento das tarefas que o processo sofreu e o processo reutilizado que não precisou de passar por todos os detalhes na sua implementação. Durante a análise qualitativa forma identificadas as tarefas da tabela 4.1 em uso, como se documenta a seguir.

TNº Prévia, para indicar a tarefa prévia e,

TNº Reutilizado, para indicar o processo reutilizado.

Exemplo: **T1 Prévia** que indica **tarefa 1 do processo prévio**.

T1 Prévia: Foi preciso inicialmente, identificar quais dados são necessários para prever o rendimento das culturas. Porém, foi fundamental a realização de um estudo para identificar os principais fatores que influenciam o rendimento das culturas. Neste estudo descobriu-se que fatores climáticos, condições de solo e humanos são os maiores responsáveis pelo rendimento das culturas. Em seguida foi necessário descobrir quais dados de cultura são necessários para prever o rendimento das culturas. Para esta seleção baseamo-nos no estudo em que se aplica o processo de DCBD no setores agrícola, pois diversos atores selecionam o mesmo tipo de variáveis para prever o rendimento das culturas. Assim sendo, escolheram-se as variáveis climáticas e/ou meteorológicas para complementar o conjunto de dados, porque, segundo De Negri e Cavalcante [De 14], para a determinação de um bom rendimento da produtividade torna-se fundamental em considerar os fatores meteorológicos, dado que estes eventos têm ficado cada vez mais rigorosos. O outro fator tomado em consideração no estudo foi a identificação da região de maior produção da cultura prévia, tendo trabalhado com cada município produtor da cultura.

T1 Reutilizado: Conhecimento de dados necessários, com base na lista disponibilizada na base de conhecimento.

T2 Prévia: Com ajuda da planilha do Excel foi possível analisar cada variável contida em cada conjunto de dados colhidos. Durante a análise foi possível verificar que os dados

que correspondem ao conjunto de dados, com informação para as culturas, possuíam um formato inadequado para o processo. Perante essa situação, cálculos com base na regra de três simples foram efetuados para cada região e para cada período correspondente.

T2 Reutilizado: Aplicação do método de três simples sobre o conjunto de dados agrícolas.

T3 Prévia: Para executar o processo de DCBD é preciso que toda a informação esteja em um mesmo conjunto de dados. Com ajuda do Excel analisamos cada variável por forma a encontrar uma variável comum que servisse como conetor entre as variáveis dos conjuntos de dados. Entre os municípios foi a variável comum dos conjuntos de dados que serviu como elo de correspondência entre eles para compor um único conjunto de dados.

T3 Reutilizado: Usando os municípios integrar os dados dos conjuntos de dados para compor um único conjunto de dados.

T4 Prévia: Devido a integração de dados pelos municípios, verificou-se que para dados meteorológicos existem poucas estações para a região, entretanto, na correspondência, alguns municípios ficaram sem dados. Para completar esta informação foi preciso interpolar os campos vazios com dados das médias dos municípios com estações mais próximas às regiões sem dados, onde encontramos estação meteorológica distante das outras. Sendo assim usou-se o mesmo valor para preencher os campos vizinhos.

T4 Reutilizado: Preencher os municípios sem estações com os dados dos valores das interpolações calculadas e pelos respectivos valores dos municípios vizinhos.

T5 Prévia: Campos com dados faltantes resultantes de falta de preenchimento ou dados desconhecidos devem ser tratados. Um caractere que não inferir nos resultados é selecionado para o preenchimento desses campos.

T5 Reutilizado: Preenchimento dos campos com falta de informação usando o caractere "?".

T6 Prévia: Analisou-se a correlação entre as variáveis que compõem o conjunto de dados para averiguar a correlação existente entre eles. Analisou-se também, o conjunto de dados com objetivo de entender o comportamento dos mesmos. Nesta sequência foi possível verificar que os dados do rendimento apresentam diferenças mensais de forma desfazada.

T6 Reutilizado: Opcionalmente pode-se analisar as variáveis do conjunto de dados.

T7 Prévia: Devido ao desfazamento verificado na análise dos dados, com ajuda do excel, agrupou-se todos os dados do conjunto nos respectivos meses ignorando os anos respectivos e, em seguida, estratificou-se esse conjunto de dados gerando 12 subconjuntos de dados para uma análise menor possível.

T7 Reutilizado: Estratificar o conjunto de dados em subconjuntos mensais ignorando os anos de produção.

T8 Prévia: Os conjuntos de dados apresentam dados com padronização diferente o que pode inferir nos resultados dos algoritmos. Testes foram executados para averiguar esta situação. Os resultados mostram-se favoráveis em conjuntos de dados normalizados.

T8 Reutilizado: Aplicar o filtro de normalização em todos os subconjuntos de dados.

T9 Prévia: Dos algoritmos estudados, foram selecionados para a tarefa de mineração de dados doze algoritmos preditores. Dos resultados obtidos alguns algoritmos mostraram-se com melhor desempenho em relação aos outros. A vantagem de executar todos algoritmos nestes subconjuntos de dados é conhecer o algoritmo com melhor desempenho para cada mês a ser previsto.

T9 Reutilizado: Aplicar os dez algoritmos preditores opcionalmente ou o algoritmo que se mostrou com melhor desempenho em relação aos outros ou ainda aplicar diretamente o algoritmo se pretender prever em um mês específico.

T10 Prévia: Foram selecionados para o estudo as métricas dos erros para análise dos algoritmos em problemas de previsão baseados em regressão. Das métricas dos erros existentes, selecionou-se dois deles a serem usados nesta dissertação. Também foi necessário analisar as variáveis contidas nos conjuntos dos dados através do coeficiente de correlação para entender o quão eles se correlacionam.

T10 Reutilizado: Analisar as métricas dos erros de predição baseada em regressão.

A partir das análises extraídas das tarefas do processo de DCBD é possível observar que as tarefas realizadas no processo prévio foram mais detalhadas em relação às tarefas do processo reutilizado. Devido a essas tarefas e configurações exercidas, o tempo gasto no processo prévio foi superior ao tempo gasto no processo reutilizado. O processo reutilizado apenas precisou aplicar as configurações contidas na base do conhecimento sem se preocupar com detalhes para a sua execução.

Em suma, foi possível obter solução do método a partir da análise qualitativa onde pode-se:

- Obter informações que auxiliem as realização do processo;
- Reagir rapidamente à execução das tarefas e etapas do processo;
- Fazer um melhor uso da base de conhecimento;
- Verificar a agilidade na execução do processo de DCBD.
- Permitir a realização dos planejamentos apropriados em curto tempo pelos responsável do setor agrícola.

Como forma de medir o ganho obtido no processo detalhado, a partir das tarefas executadas, analisaremos quantitativamente o processo, testando assim a viabilidade para a solução proposta.

4.2 ANÁLISE QUANTITATIVA DO PROCESSO

Depois de analisar qualitativamente as tarefas exercidas no processo prévio e reutilizado, com base no obtido, uma análise qualitativa vai aferir o resultado para medir quanto de ganho de tempo a solução teve na duração da execução do processo. Para tal, listamos a duração de cada tarefa dos dois processos executados, como podemos ver na tabela 4.2.

Tabela 4.2: Experiência para comprovar a viabilidade

Tarefa	**Ferramenta**	**Duração do processo inicial (≈)**	**Duração do segundo processo (≈)**
01	Fontes de buscas	3 dias	1 dia
02	Excel	2 dias	1 dia
03	Excel	2 dias	1 dia
04	Excel	30 dias	10 dias
05	Excel	1 dia	1 dia
06	Rattle	1 dia	1 dia
07	Excel	2 dias	1 dia
08	Weka	1 dia	1 dia
09	Weka	12 dias	12 dias
10	Weka	5 dias	5 dias

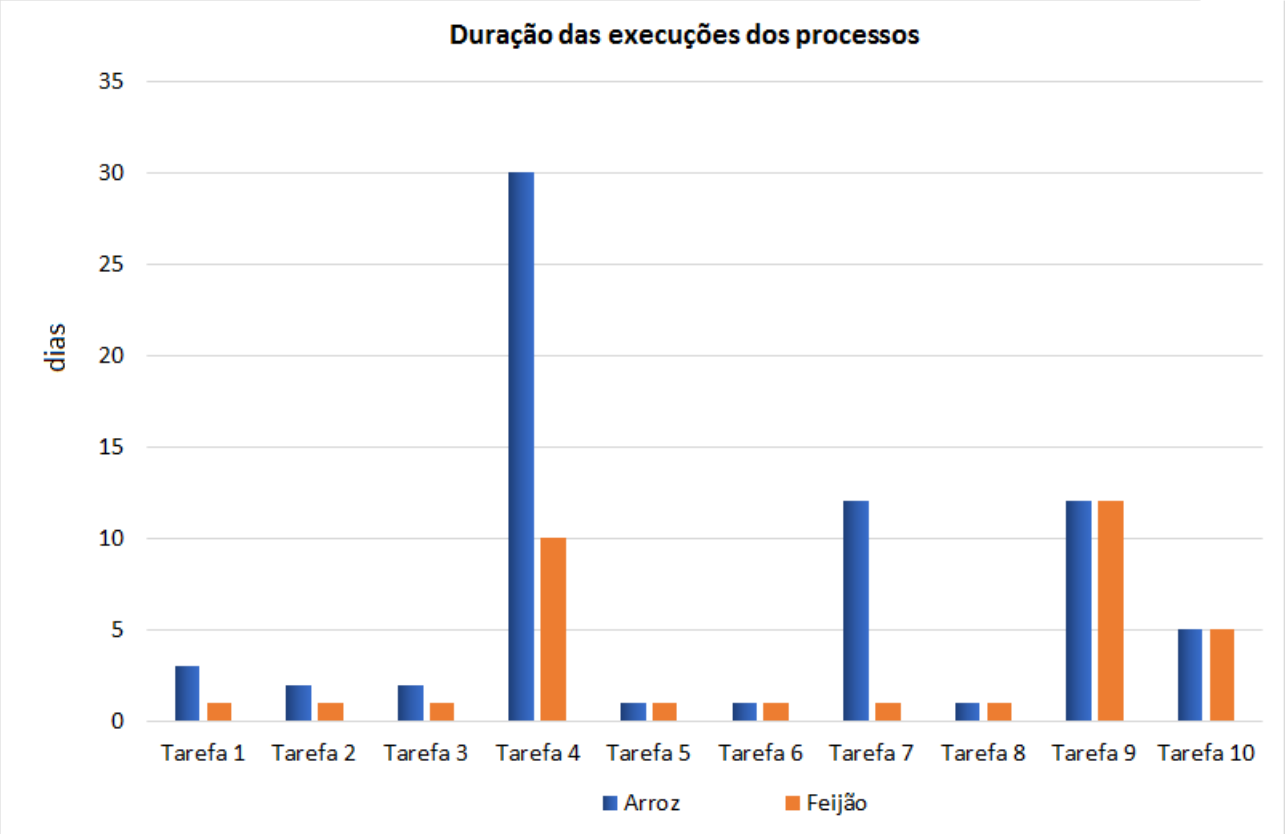


Figura 4.1: Duração das tarefas do processo

Pelo apresentado na tabela, existe uma diferença substancial no tempo que é despendido na criação das tarefas configuradas sem a utilização da base de conhecimento e o tempo que é despendido com a utilização do recurso da solução desenvolvida. Esta permite identificar quais atributos devem ser selecionados e quais tarefas devem ser realizadas e como elas devem ser aplicadas na execução do novo processo.

Outra forma apurada para análise do processo, foi a comparação da duração da execução de todo o processo realizado sobre a cultura do arroz e a cultura do feijão, apresentados na figura 4.1. Feitas as medidas de forma porcentual da execução das fases do processo de DCBD, foi possível apurar, a partir dos resultados obtidos, que com o uso da base de conhecimento aprendido, usado como guia para a execução de um processo de DCBD no setor agrícola, torna-se viável para a agilização do novo processo.

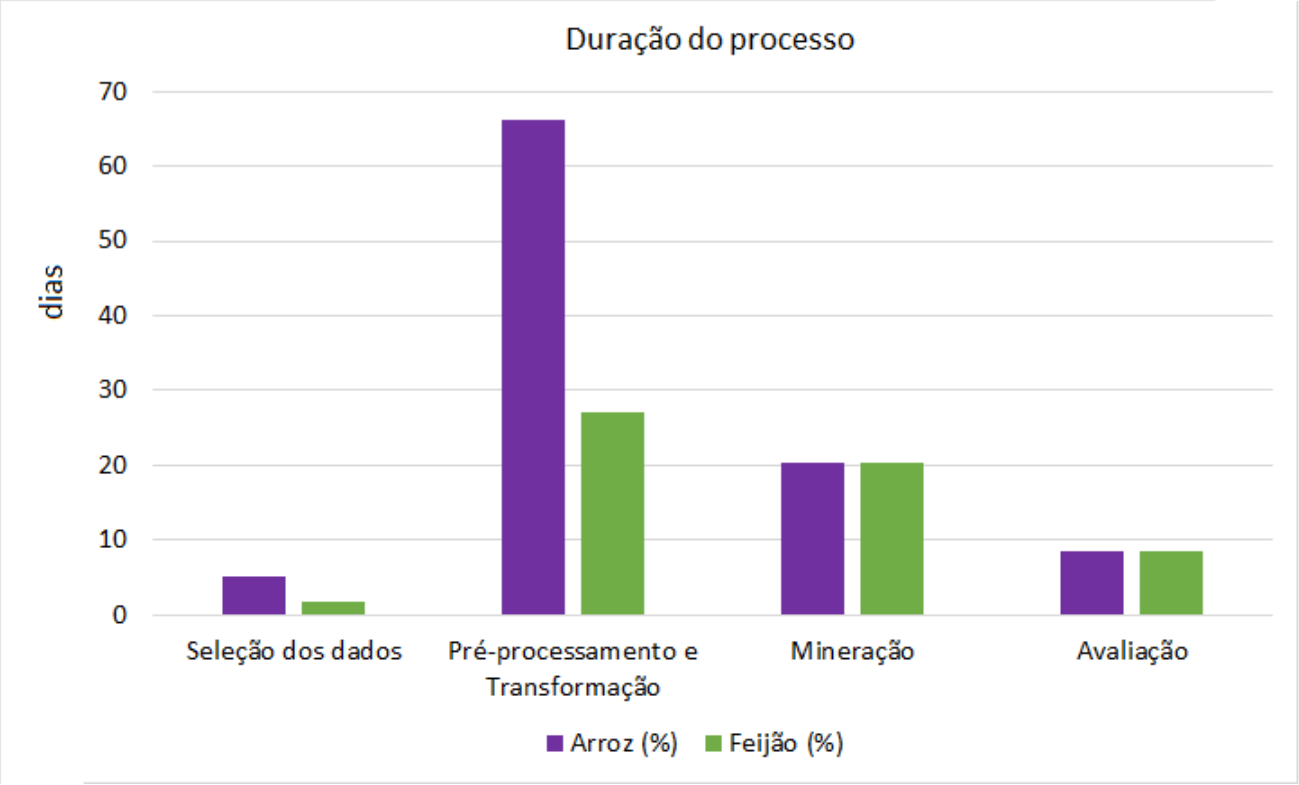


Figura 4.2: Duração do processo por fases

Na figura 4.2 podemos verificar que a fase de seleção dos dados, pré-processamento e transformação são as fases que ganharam agilidade na sua execução, isto porque estas são as fases mais essências de todo o processo. Pelos resultados atingidos, apresentados na tabela 4.3, podemos concluir que o nosso estudo atingiu o objetivo definido em que o modelo reutilizado teve um ganho relativo em 42,37% de agilidade em relação ao tempo de execução total do processo de DCBD reutilizado.

Tabela 4.3: Duração das execuções dos processos por fases

Fases	**Arroz (dias)**	**Arroz (%)**	**Feijão (dias)**	**Feijão (%)**	**Ganho relativo (%)**
Seleção dos dados	3	5,08	1	2,9	66,66
Pré-processamento e transformação	39	66,10	16	47,1	58,97
Mineração dos dados	12	20,34	12	35,3	0
Avaliação	5	8,47	5	14,7	0
Total	59	100	34	57,63	42,37
Ganho do processo	**42,37%**				

5. CONCLUSÃO

O setor agrícola está sujeito, desde sempre, à manipulações tendentes a aquisição de métodos eficientes e eficazes para a previsão de rendimento das culturas. Entre as várias formas de descobertas de conhecimento em base de dados para a previsão do rendimento, inquietou-nos a componente da falta e/ou o desconhecimento de reutilização do processo de descoberta de conhecimento em base de dados agrícolas, que culmina com a perda de tempo na repetição das tarefas executadas.

Com base no exposto, fez-se uma análise do processo executado na cultura do arroz. As tarefas bem sucedidas nessa cultura permitiram a criação de uma base de conhecimento resultante desse processo. Essa base foi reutilizada na cultura do feijão. Os resultados mostraram uma agilidade na execução do processo de DCBD para o setor agrícola.

A aplicação da base de conhecimento do processo fonte em novos processos, faz com que as próximas execuções em culturas temporárias sejam agilizadas. Assim, por parte dos especialistas/analistas de domínio, esforços são evindados durante a execução do processo pelo conhecimento de configurações necessárias para a execução das tarefas do processo.

Quando se tenta aplicar o mesmo processo sobre uma nova fonte de dados agrícola, existe a necessidade de readaptação do processo de descoberta de conhecimento desenvolvido. No entanto, é possível afirmar que depois que executou-se o processo da cultura do feijão, nem todas as fases precisam de sofrer uma reestruturação/adaptação, mas que estas podem ser agilizadas quando é conhecido o processo sobre o qual deve ser aplicado.

Uma vez que o objetivo é aplicar as mesmas técnicas de processamento e mineração de dados que tinham sido aplicadas sobre a primeira fonte (cultura do arroz), o que é necessário é refazer as primeiras fases de DCBD e aproveitar assim as restantes fases definidas do processo anterior. Para tal, deve-se reutilizar o modelo implementado no processo do arroz desde a fase da seleção de dados até aos modelos definidos na fase de mineração de dados e deverá ser respeitado sempre que este processo seja reaplicado a um novo conjunto de dados (uma nova cultura).

O objetivo desta dissertação é criar uma metodologia que permita reutilizar o processo de DCBD, otimizando assim as tarefas desenvolvidas pelos analistas, para que estes não despendam tempo na configuração e na transformação dos dados, rentabilizando o tempo na extração de conhecimento ou em outras atividades. Desta forma, esta metodologia irá abranger não só a preparação dos dados, como também o pré-processamento dos mesmos, permitindo que possam ser executados os algoritmos de mineração de dados,

recorrendo apenas a ajustes quando aplicado sobre uma nova base de dados no domínio das culturas temporárias.

Descreveu-se esta abordagem para facilitar a integração baseada em padrões e parâmetros do processo de DCBD. Demonstramos que esses parâmetros permitem uma reutilização fácil que podem acelerar significativamente o processo da descoberta de conhecimento em base de dados agrícola, onde avaliamos a abordagem em um estudo de caso para a cultura do feijão no Estado do Rio Grande do Sul.

Foi possível durante a execução do método, identificar parâmetros essenciais que permitem a reutilização do processo. Identificou-se os tipos de dados necessários para o processo executado em um domínio de culturas temporárias dos quais dados meteorológicos, dados das coordenadas geográficas e dados da cultura. Estes dados sofreram configurações durante a sua modelagem no processo e foram catalogadas permitindo assim o fácil aproveitamento em novos processos.

Testou-se o modelo de reutilização do processo DCBD com base no conhecimento da execução na cultura do arroz. Os resultados foram o esperado, pois o processo realizado no feijão foi mais ágil em relação a execução ao processo do arroz em 42,37% do ganho de tempo de execução devido às configurações aproveitadas da execução do arroz que facilitaram todo o processo de DCBD executado. Salientar que este processo foi testado para aplicar apenas em culturas temporárias.

5.1 CONTRIBUIÇÃO

Neste trabalho sobre a reestruturação e adaptação, que define a configuração criada de um processo de descoberta de conhecimento em dados agrícolas para a previsão do rendimento das culturas, foi implementada uma metodologia sobre como seria possível flexibilizar a execução do processo por meio da reutilização de um processo de descoberta de conhecimento previamente definido inicialmente, tendo em conta a necessidade inicial de reduzir o tempo despendido.

Os objetivos propostos inicialmente foram atingidos, tendo-se criado a metodologia com a solução proposta com base em conhecimento aprendido da execução inicial do processo, bem como de outros problemas detetados com o decorrer do trabalho.

A criação de uma nova metodologia era um dos objetivos propostos. Essa metodologia foi alcançada. Esta permite reutilizar um processo de descoberta de conhecimento em dados agrícolas partindo da sua configuração inicial executada sobre a cultura do arroz, reaproveitando ações que o analista efetuou na criação desse processo.

Com a criação dessa metodologia, foi possível alcançar e criar uma base de conhecimento que dispensa menos tempo na reconfiguração/adaptação das tarefas do pro-

cesso. Para tal, foi criada uma lista de aprendizado com a parametrização necessária dos passos definidos pelo método criada, readaptando assim o processo de DCBD utilizado.

Antes de finalizar, importa sublinhar que a visão colocada neste trabalho não só pode revolucionar a atividade agrícola como também pode contribuir na simplificação das atividades desenvolvidas pelos especialistas/analistas de dados. Mas, sobretudo, na aplicação dessa estratégia em contexto moçambicano - a origem da autora desta dissertação.

5.2 LIMITAÇÕES

Uma das maiores limitações encontradas foi o curto tempo de pesquisa que não corroborou com a pouca execução de testes para validar a solução, mas no entanto, foi possível deixar a experiência testada em apenas uma cultura.

Outra limitação que foi ultrapassada, foi que durante a execução do processo prévio inicial, deparou-se com dificuldades para encontrar as estações meteorológicas vizinhas. Esta limitação foi solucionada com a realização do cálculo da distância usando dados das coordenadas geográficas entre os municípios produtores em relação aos municípios com estações meteorológicas, que não foi um trabalho fácil de se realizar.

5.3 LIÇÕES APRENDIDAS

Resume-se as lições aprendidas dos resultados das execuções do processo de DCBD sobre a cultura do arroz. Deixa-se aqui algumas condições que acredita-se serem cumpridas em um processo de DCBD bem-sucedido na agricultura:

- **Conhecer o domínio inserido:** Estando em um problema específico é suposto que este problema seja de conhecimento do usuário. No caso, a necessidade de envolver o especialista em domínio no processo demorado e tedioso de obter conhecimento acaba sendo minimizado.

  No entanto, no processo de DCBD agrícola e no mundo real, as fases mais difíceis são a compreensão de dados e a preparação dos mesmos. A fase de preparação de dados para além de ser mais difícil é a fase mais demorada de todo o processo, nele é gasto cerca de 66,10% do tempo de processamento.

- **Selecionar dados externos:** Há muitos fatores externos que não são coletados diretamente para a tarefa de DCBD, mas pode ter um grande impacto na análise de dados. Dados provenientes de fontes diferentes impactam na análise de dados e sobrecarregam a tarefa de pré-processamento dos mesmos dados. Deve-se selecionar

dados relacionados com a produção agrícola, dados meteorológicos e coordenadas geográficas da localidade em estudo.

- **Usar métodos de pré-processamento e transformação:** As ações típicas de pré-processamento são tabelas de junção (integração), agregação, normalização, manipulação de valores ausentes, criação de novos atributos. Muitas vezes, essas operações são realizadas como domínio independente, o que faz quando num processo diferente, mas do mesmo domínio, repitam todas as tarefas - uma prática que pode ser evitada se deixando tudo configurado para que seja reutilizado nos próximos processos.
- **Usar modelos simples de regressão** As vezes os relatórios e resumos “simples” geram resultados aceitáveis em vez de procurar usar algoritmos sofisticados e não simplesmente buscar resultados simples para o problema. Uma vez conhecido os algoritmos que melhor apresentam o seu desempenho, evita-se vários testes na sua execução, o que pode ser minimizado através da execução do algoritmo conhecido e testado em outras situações, ganhando assim a agilidade no processo de DCBD.

5.4 TRABALHOS FUTUROS

Como trabalhos futuros espera-se:

- Implementar um sistema baseada em reutilização do processo de DCBD agrícola. O objetivo deste sistema poderá executar o processo de forma ágil a partir das configurações efetuadas no próprio sistema. O sistema poderá prover um guia de tarefas a serem efetuadas para atingir o objetivo definido.
- Executar mais experimentos usando culturas de regiões diferentes para testar a aplicabilidade do método reutilizável.
- Testar o método com dados reais de um setor agrícola de moçambique de forma a ajudar o setor a planificar as suas atividades dentro do tempo.
- Interpretar o modelo dos algoritmos com melhor desempenho para obter o conhecimento necessário que traduz o resultado da previsão do rendimento das culturas.
- Produzir e publicar artigos relacionados ao assunto por forma a contribuir cientificamente com a ideia desenvolvida.

REFERÊNCIAS BIBLIOGRÁFICAS

[Aze19] Azevedo, A. “Data mining and knowledge discovery in databases”. In: *Proceedings of the Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*, IGI Global, 2019, pp. 502–514.

[Bal19] Balbino, A. “O impacto das mudanças climáticas na agricultura”. Capturado em: https://agrosmart.com.br/blog/impacto-mudancas-climaticas-na-agricultura/, Agosto 2019.

[BB18] Bhojani, S. H.; Bhatt, D. N. “Application of data mining technique for wheat crop yield forecasting for districts of gujarat state”, *International Journal of Scientific and Research Publications*, vol. 8–7, Julho 2018, pp. 302–306.

[BGE08] Boente, A. N. P.; Goldschmidt, R. R.; Estrela, V. V. “Uma metodologia de suporte ao processo de descoberta de conhecimento em bases de dados”. In: Anais do V Simposio de Excelencia em Gestão e Tecnologia, 2008, pp. 4–5.

[BM14] Basha, N.; Mohan, C. “A methodology to identify the level of reuse using template factors”, *arXiv preprint*, vol. 1406.3727, 2014, pp. 103–114.

[Car15] Carmo, C. R. S. “Culturas temporárias no brasil: um estudo sobre possíveis determinantes da área cultivada ao longo dos anos 1991 a 2012”, *Revista GeTeC*, vol. 4–7, Janeiro-Junho 2015, pp. 55–78.

[CHL06] Cano, J. R.; Herrera, F.; Lozano, M. “On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining”, *Applied Soft Computing*, vol. 6–3, Março 2006, pp. 323–332.

[DA-18] DA-IF/UFRGS, D. d. A. d. I. d. F. d. U. “Coordenadas geográficas”. Capturado em: http://astro.if.ufrgs.br/br.htm/&https://pt.db-city.com/Brasil, Abril 2018.

[DB13] Diriba, Z.; Borena, B. “Application of data mining techniques for crop productivity prediction”, *HiLCoE Journal of Computer Science and Technology*, vol. 1–2, Junho 2013, pp. 152–155.

[De 14] De Negri, Fernanda e Cavalcante, L. R. “Produtividade no brasil: desempenho e determinantes. brasília: Abdi”, vol. 1, Novembro 2014, pp. 373–409.

[Del19] Delfino, C. “Calculando distâncias com base em coordenadas de gps”. Capturado em: http://carlosdelfino.eti.br/cursoarduino/geoprocessamento/calculando-distancias-com-base-em-coordenadas-de-gps/, Maio 2019.

[dGeE18] de Geografia e Estatística, I. B. “Levantamento sistemático da produção agrícola - lspa”. Capturado em: https://www.embrapa.br/agropensa/producao-agricola-municipal/, Abril 2018.

[dGeEI19] de Geografia e Estatística (IBGE), I. B. “Pesquisa agrícola municipal”. Capturado em: https://www.brasil247.com/pt/247/rs247/118839/RS-é-o-maior-produtor-de-arroz-do-país-aponta-IBGE.htm/, Maio 2019.

[DMU17] Dey, U. K.; Masud, A. H.; Uddin, M. N. “Rice yield prediction model using data mining”. In: Proceedings of the International Conference on Electrical, Computer and Communication Engineering, 2017, pp. 321–326.

[dRGdS19] do Rio Grande do Sul, A. S. “Clima, temperatura e precipitação”. Capturado em: https://atlassocioeconomico.rs.gov.br/clima-temperatura-e-precipitacao, Agosto 2019.

[dSSdL+16] dos SANTOS, B. S.; STEINER, M. T. A.; de Lara, L. H.; MARTINS, L. G. R.; ANDRADE, d. L.; Rochavetz, P. “Data mining: Uma abordagem teórica e suas aplicações”, *Revista Espacios*, vol. 37–05, Novembro 2016, pp. 23–23.

[ESEMRE13] El-Sappagh, S.; El-Masri, S.; Riad, A.; Elmogy, M. “Data mining and knowledge discovery: applications, techniques, challenges and process models in healthcare”, *International Journal of Engineering Research and Applications*, vol. 3–3, Abril 2013, pp. 900–906.

[FEBL09] Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. “Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses”, *Behavior research methods*, vol. 41–4, Novembro 2009, pp. 1149–1160.

[FLG+11] Faceli, K.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. d. L.; et al.. “Inteligência Artificial: Uma abordagem de aprendizado de máquina”. LTC, 2011.

[FMZ15] Fetanat, H.; Mortazavifarr, L.; Zarshenas, N. “The analysis of agricultural data with regression data mining technique”, *Ciência e Natura*, vol. 37–6-2, Junho 2015, pp. 102–107.

[FPSS96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. “From data mining to knowledge discovery in databases”, *AI Magazine*, vol. 17, Setembro 1996, pp. 37–54.

[GCJ15] Garcia, E.; Camolesi Jr, L. “Aplicação do processo de descoberta de conhecimento em base de dados agrícola para reconhecimento de fatores que impactam na produtividade da cana-de-açúcar”. In: Anais do XI Congresso Nacional de Excelência em Gestão, 2015, pp. 415–427.

[GP05] Goldschmidt, R.; Passos, E. “Data mining: um guia prático”. Gulf Professional Publishing, 2005.

[GS18] Geetha, M.; Shanthi, I. E. “Predicting the soil profile through modified regression by discretisation algorithm for the crop yield in trichy district, india”, *International Journal of Grid and Utility Computing*, vol. 9–3, Julho 2018, pp. 235–242.

[HPK11] Han, J.; Pei, J.; Kamber, M. “Data mining: concepts and techniques”. Elsevier, 2011.

[J+13] Jagtap, S. B.; et al.. “Census data mining and data analysis using weka”, *ArXiv Preprint*, vol. 1310.4647, Outubro 2013, pp. 35–40.

[KI17] Kodeeshwari, R.; Ilakkiya, K. T. “Different types of data mining techniques used in agriculture-a survey”, *International Journal of Advanced Engineering Research and Science*, vol. 4–6, Junho 2017, pp. 17–23.

[KKPB17] Kamilaris, A.; Kartakoullis, A.; Prenafeta-Boldú, F. X. “A review on the practice of big data analysis in agriculture”, *Computers and Electronics in Agriculture*, vol. 143, Dezembro 2017, pp. 23–37.

[KS13] Kesavaraj, G.; Sukumaran, S. “A study on classification techniques in data mining”. In: Proceedings oh the 4th International Conference on Computing, Communications and Networking Technologies, 2013, pp. 1–7.

[Len02] Lenzerini, M. “Data integration: A theoretical perspective”. In: Proceedings of the 21st ACM Special Interest Group on Management of Data - Special Interest Group for Automata and Computability Theory - Special Interest Group for Artificial Intelligence Tutorial Symposium on Principles of Database Systems, 2002, pp. 233–246.

[Met18] de Meteorologia, I. I. N. “Dados históricos de meteorologia”. Capturado em: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep, Agosto 2018.

[MHN15] Moor, H.; Hylander, K.; Norberg, J. “Predicting climate change effects on wetland ecosystem services using species distribution modeling and plant functional traits”, *Ambio*, vol. 44–1, Janeiro 2015, pp. 113–126.

[MNA17] Majumdar, J.; Naraseeyappa, S.; Ankalaki, S. “Analysis of agriculture data using data mining techniques: application of big data”, *Journal of Big Data*, vol. 4–1, Dezembro 2017, pp. 1–15.

[MPdL+17] Machado, J.; Padilha, M. d. R. d. F.; de Lira, F. P.; de Oliveira, J. G.; da Silva, R. S.; Caetano, M. B. C. “Agricultura de precisão e abertura de novas fronteiras no brasil| precision agriculture and opening new frontiers in brazil”, *Revista Geama*, vol. 4–1, Dezembro 2017, pp. 49–53.

[MR11] Mucherino, A.; Ruß, G. “Recent developments in data mining and agriculture.” In: Proceedings of the Industrial Conference on Data Mining-Workshops, 2011, pp. 90–98.

[PGCMA15] Pacheco, C. L.; Garcia, I. A.; Calvo-Manzano, J. A.; Arcilla, M. “A proposed model for reuse of software requirements in requirements catalog”, *Journal of Software: Evolution and Process*, vol. 27–1, 2015, pp. 1–21.

[PGM+14] Pereira, V. G. C.; Gris, D. J.; Marangoni, T.; Frigo, J. P.; de Azevedo, K. D.; Grzesiuck, A. E. “Exigências agroclimáticas para a cultura do feijão (phaseolus vulgaris l.)”, *Revista Brasileira de Energias Renováveis*, vol. 3–1, 1º trimestre 2014, pp. 32–42.

[PRR+17] Pudumalar, S.; Ramanujam, E.; Rajashree, R. H.; Kavya, C.; Kiruthika, T.; Nisha, J. “Crop recommendation system for precision agriculture”. In: Proceedings of the 8th International Conference on Advanced Computing, 2017, pp. 32–36.

[PT14] Priyadharsini, C.; Thanamani, A. S. “An overview of knowledge discovery database and data mining techniques”, *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2–1, 2014, pp. 1571–1578.

[RK10] Ruß, G.; Kruse, R. “Regression models for spatial data: An example from precision agriculture”. In: Proceedings of the Industrial Conference on Data Mining, 2010, pp. 450–463.

[RWB10] Rüping, S.; Wegener, D.; Bremer, P. “Re-using data mining workflows”. In: Proceedings of the 3rd-Generation Data Mining: Towards Service-Oriented Knowledge Discovery, 2010, pp. 25–30.

[SA18] Surya, P.; Aroquiaraj, I. L. “Crop yield prediction in agriculture using data mining predictive analytic techniques”, *International Journal of Research and Analytical Reviews*, vol. 5–4, Dezembro 2018, pp. 783–787.

[SAS+14a] Sanchez, G.; Alberto; Solis, F.; Juan; Bustamante, O.; Waldo. “Attribute selection impact on linear and nonlinear regression models for crop yield prediction”, *The Scientific World Journal*, vol. 2014–509429, Maio 2014, pp. 1–10.

[SAS+14b] Sánchez, G.; Alberto; Solís, F.; Juan; Bustamante, O.; Waldo; et al.. “Predictive ability of machine learning methods for massive crop yield prediction”, *Spanish Journal of Agricultural Research*, vol. 12–2, Abril 2014, pp. 313–328.

[SP15] Shepherd, C.; Palmer, L. “The modern origins of traditional agriculture”, *Bijdragen tot de taal-, land-en volkenkunde/Journal of the Humanities and Social Sciences of Southeast Asia*, vol. 171–2-3, Janeiro 2015, pp. 281–311.

[SP16] Sellam, V.; Poovammal, E. “Prediction of crop yield using regression analysis”, *Indian Journal of Science and Technology*, vol. 9–38, Outubro 2016, pp. 1–5.

[Sri14] Srivastava, S. “Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining”, *International Journal of Computer Applications*, vol. 88–10, Fevereiro 2014, pp. 26–29.

[VMS11] Veenadhari, S.; Misra, B.; Singh, C. “Data mining techniques for predicting crop productivity—a review article”, *International Journal of Computer Science and Technology*, vol. 2–1, Março 2011, pp. 90–100.

[WFHP16] Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. “Data Mining: Practical machine learning tools and techniques”. Morgan Kaufmann, 2016.

[WR10] Wegener, D.; Rüping, S. “On reusing data mining in business processes-a pattern-based approach”. In: Proceedings of the International Conference on Business Process Management, 2010, pp. 264–276.

[XG18] Xu, Y.; Goodacre, R. “On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning”, *Journal of Analysis and Testing*, vol. 2–3, Outubro 2018, pp. 249–262.

[ZWZ03] Zhang, S.; Wu, X.; Zhang, C. “Multi-database mining”, vol. 2–1, Junho 2003, pp. 5–13.

APÊNDICE A – GRÁFICO DAS CORRELAÇÕES MENSAIS DA CULTURA DO ARROZ

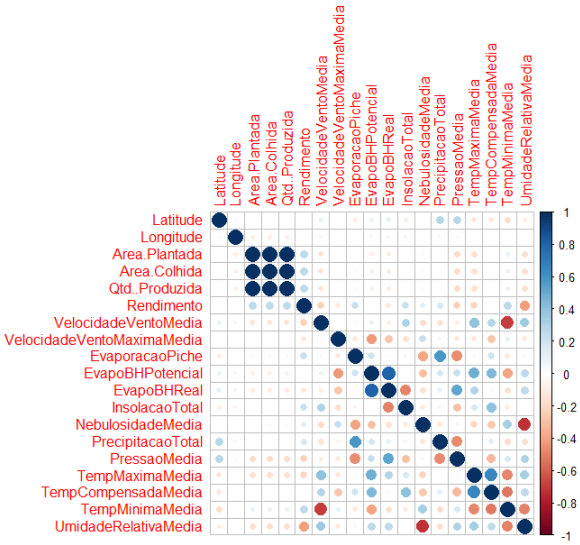


Figura A.1: Correlação de março

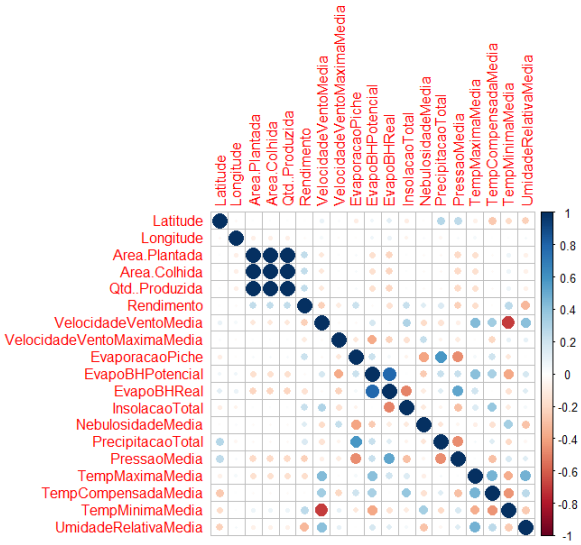
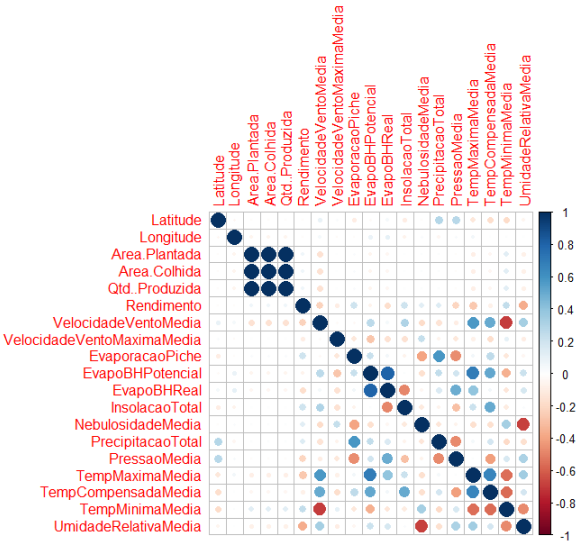


Figura A.2: Correlação de abril



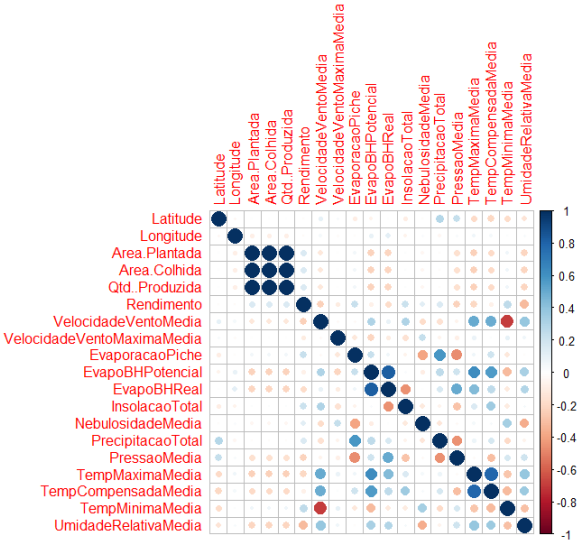


Figura A.3: Gráfico das correlações de maio e junho

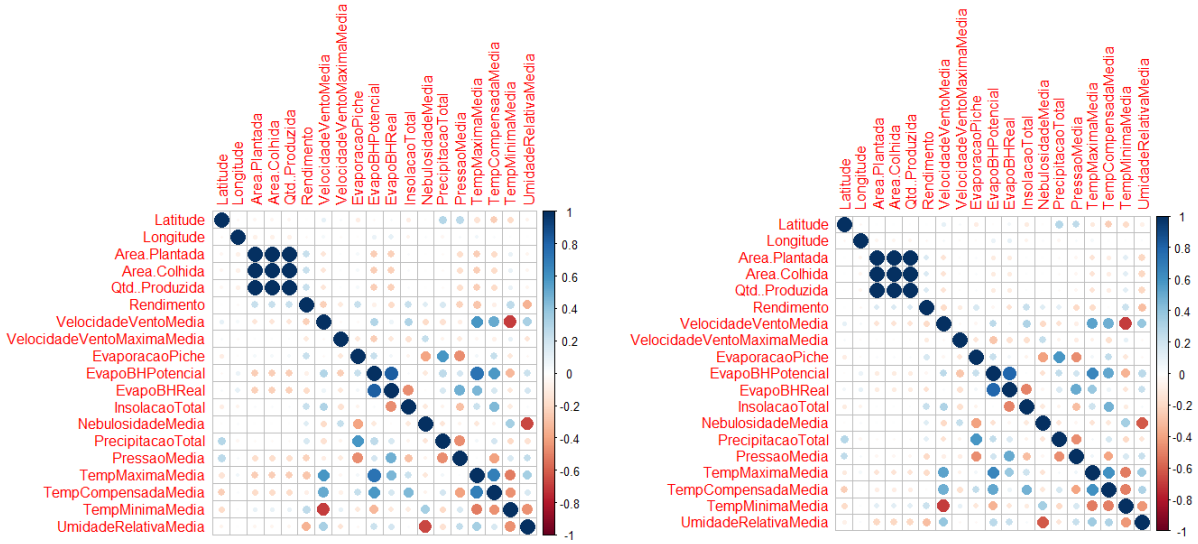


Figura A.4: Gráfico das correlações de julho e agosto

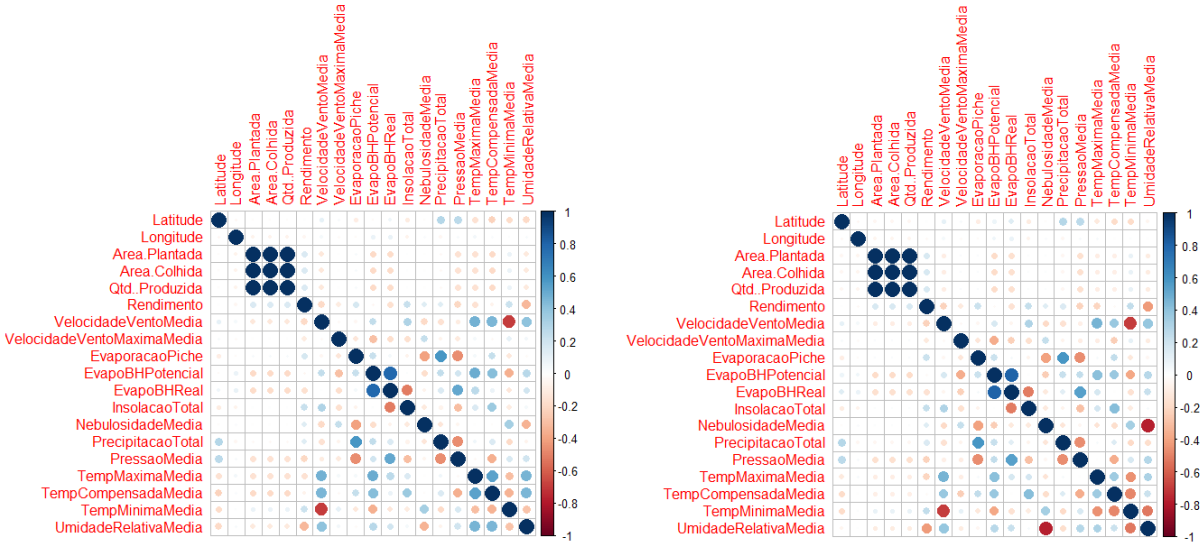


Figura A.5: Gráfico das correlações de setembro e outubro

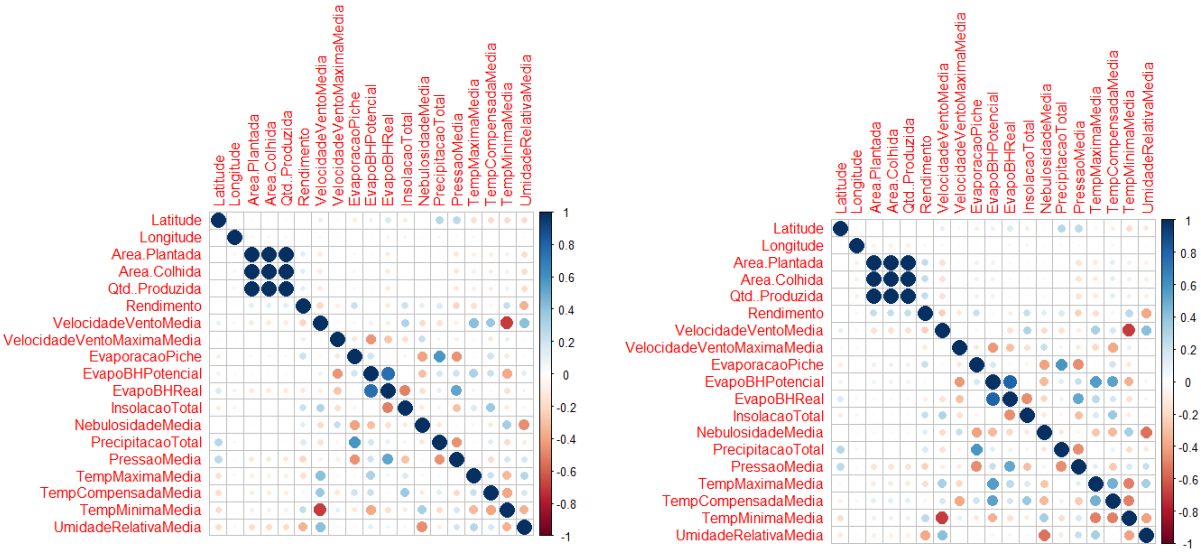


Figura A.6: Gráfico das correlações de novembro e dezembro

APÊNDICE B – GRÁFICOS DO COMPORTAMENTO MENSAL DO RENDIMENTO DA CULTURA DO ARROZ

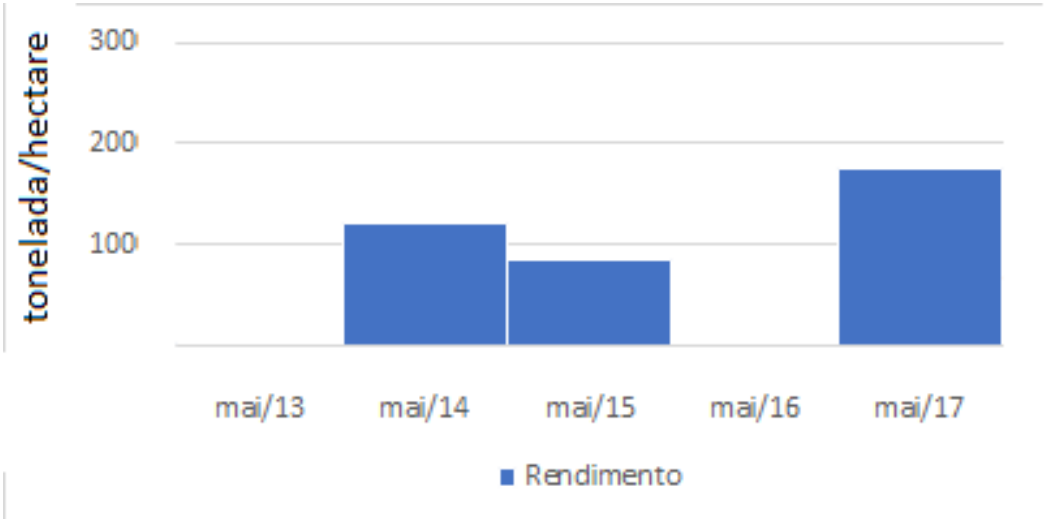


Figura B.1: Comportamento de maio

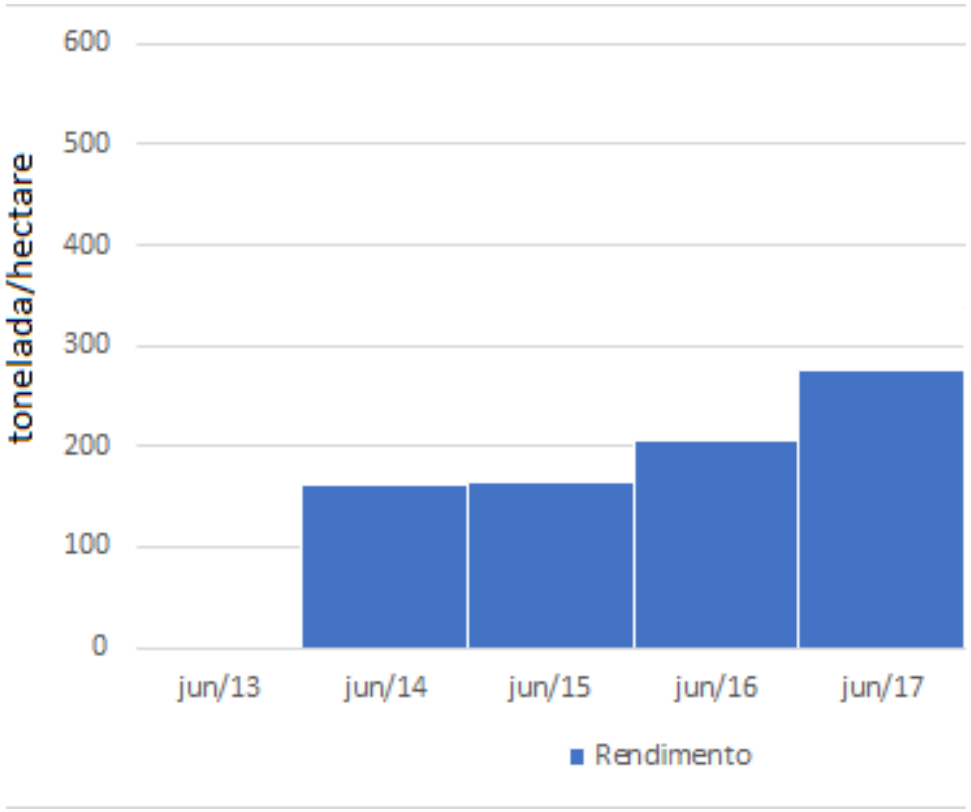


Figura B.2: Comportamento de junho

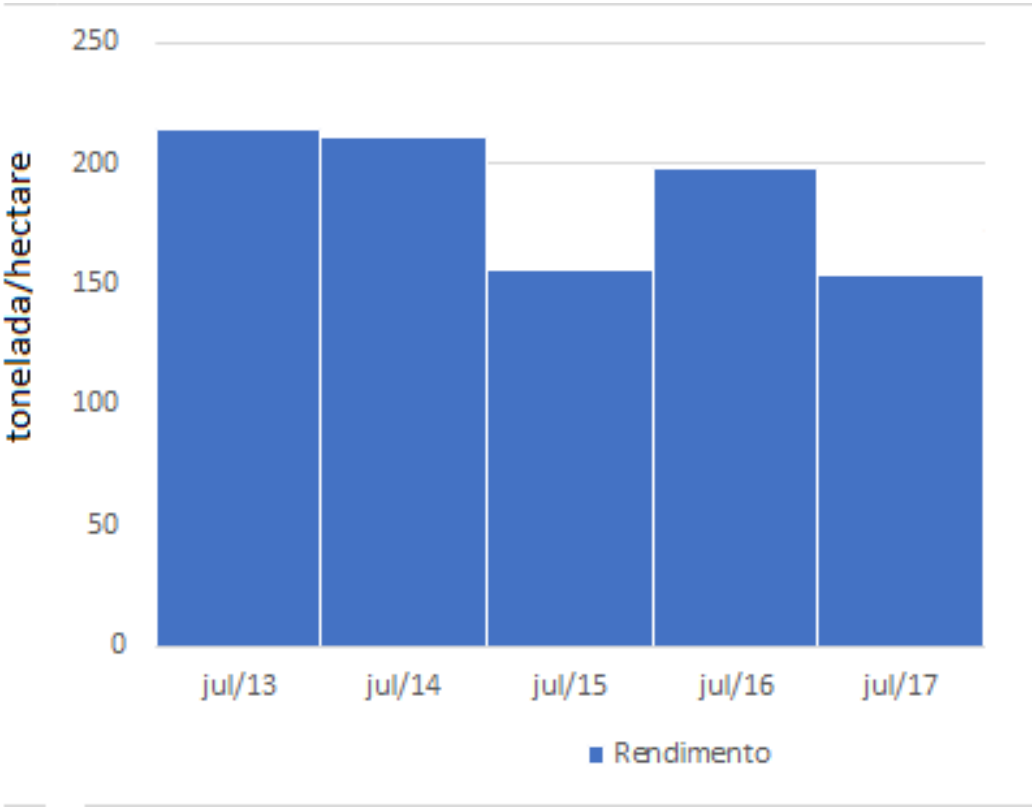


Figura B.3: Comportamento de julho

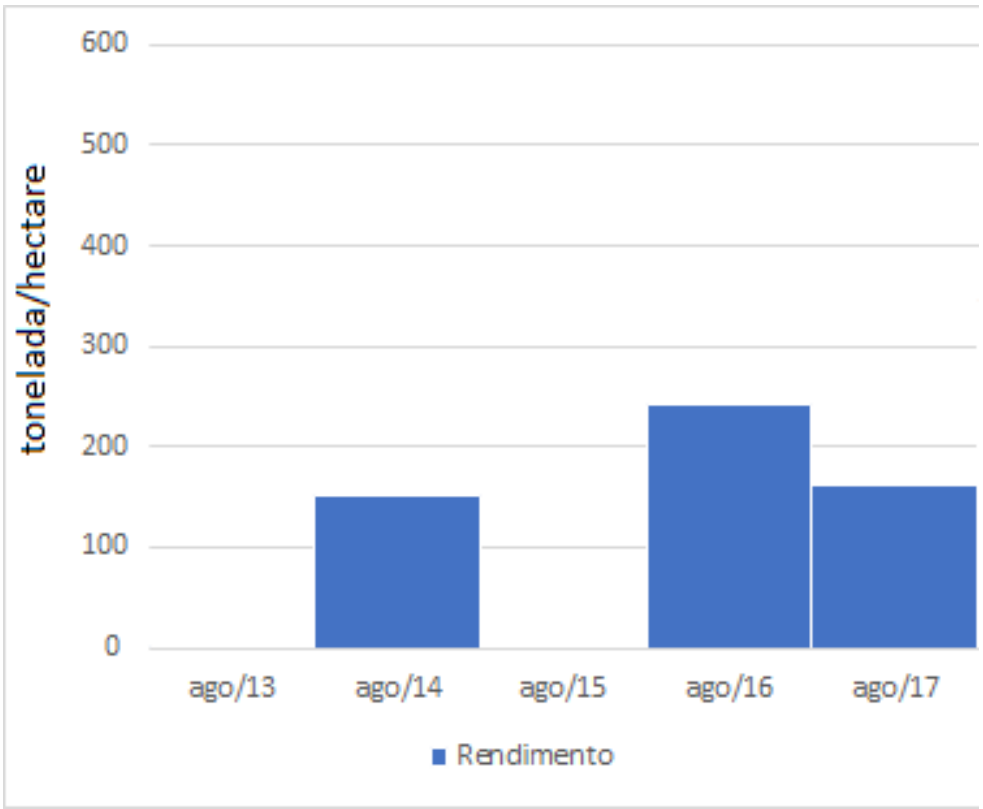


Figura B.4: Comportamento de agosto

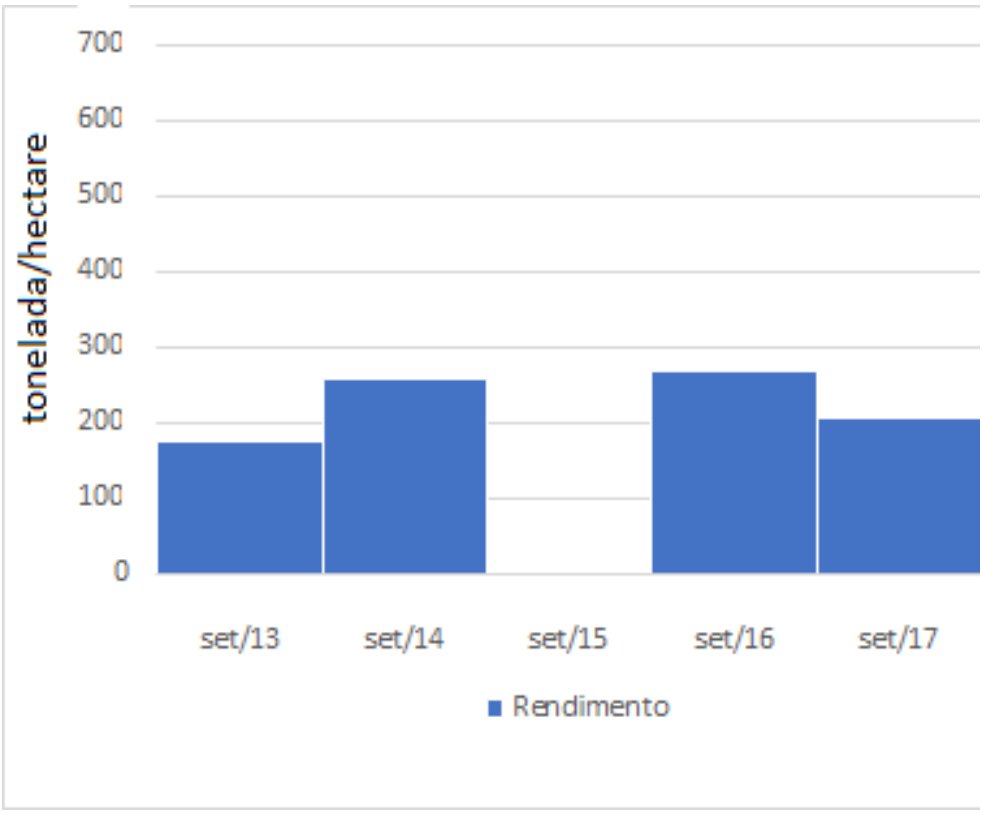


Figura B.5: Comportamento de setembro

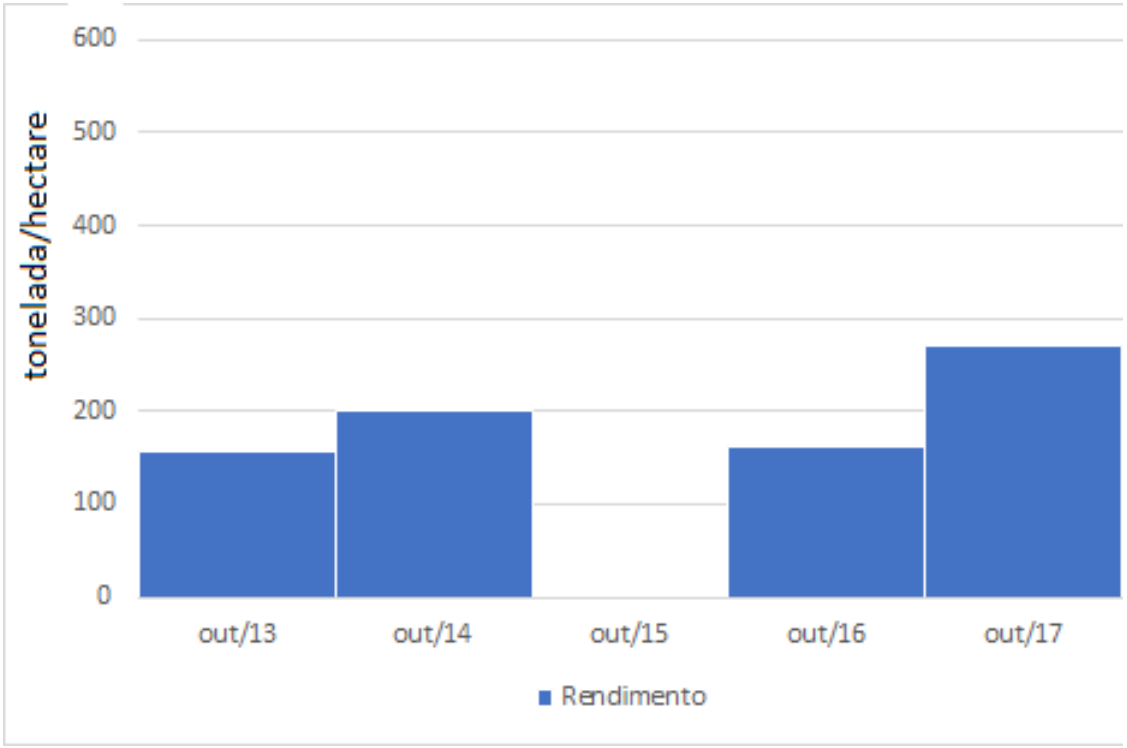


Figura B.6: Comportamento de outubro

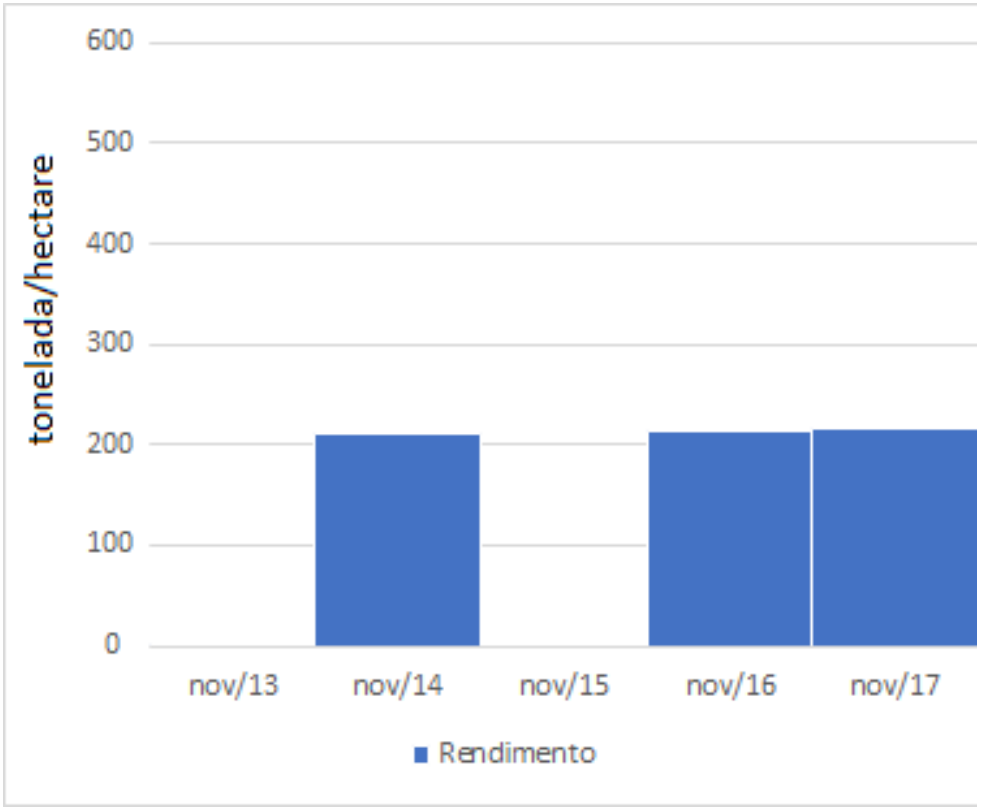


Figura B.7: Comportamento de novembro

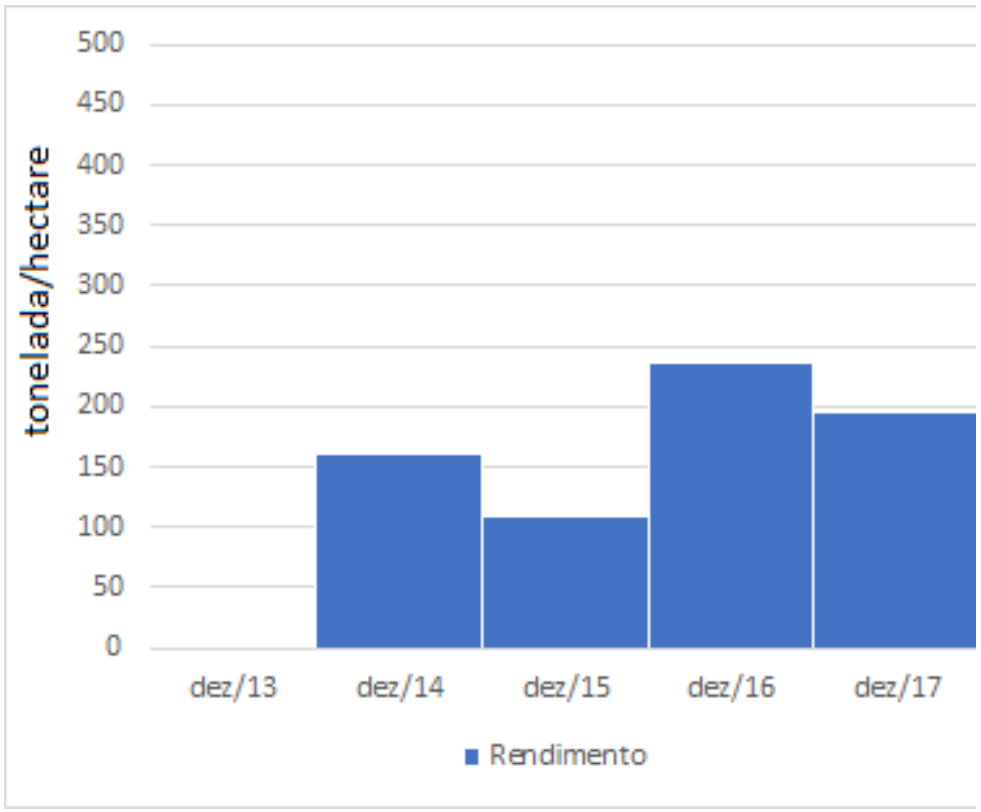


Figura B.8: Comportamento de dezembro

APÊNDICE C – TABELAS MENSAIS DA CULTURA DO ARROZ

Tabela C.1: Métrica do mês de Abril

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,932	0,0666	0,094
trees.RandomTree	0,9257	0,065	0,0983
trees.RandomForest	0,9567	0,0565	0,0763
trees.M5P	0,9536	0,0498	0,078
functions.LinearRegression	0,939	0,0618	0,0892
functions.MultilayerPerceptron	0,0011	0,2782	0,35
lazy.IBk	0,9124	0,068	0,1084
meta.AdditiveRegression	0,9208	0,071	0,1007
meta.Bagging	0,9225	0,0752	0,1007
meta.RegressionByDiscretization	0,9263	0,0653	0,0973

Tabela C.2: Métrica do mês de Maio

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,8117	0,1142	0,1506
trees.RandomTree	0,8628	0,0921	0,1331
trees.RandomForest	0,9109	0,0804	0,1055
trees.M5P	0,9439	0,0571	0,0837
functions.LinearRegression	0,855	0,1019	0,1333
functions.MultilayerPerceptron	0,0096	0,2608	0,3199
lazy.IBk	0,8547	0,1018	0,1429
meta.AdditiveRegression	0,8969	0,0801	0,1121
meta.Bagging	0,8023	0,1189	0,1516
meta.RegressionByDiscretization	0,9036	0,0745	0,1091

Tabela C.3: Métrica do mês de Junho

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,8601	0,0915	0,1234
trees.RandomTree	0,8696	0,0911	0,12
trees.RandomForest	0,9214	0,0702	0,0938
trees.M5P	0,9379	0,0551	0,0828
functions.LinearRegression	0,8891	0,0859	0,1107
functions.MultilayerPerceptron	0,1339	0,2338	0,2951
lazy.IBk	0,8406	0,0964	0,1332
meta.AdditiveRegression	0,8968	0,0713	0,1057
meta.Bagging	0,8507	0,0947	0,1255
meta.RegressionByDiscretization	0,8936	0,0747	0,1073

Tabela C.4: Métrica do mês de Julho

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,9275	0,0683	0,0956
trees.RandomTree	0,9021	0,0743	0,1121
trees.RandomForest	0,948	0,0605	0,0816
trees.M5P	0,9477	0,051	0,0813
functions.LinearRegression	0,9314	0,0669	0,0931
functions.MultilayerPerceptron	-0,026	0,2984	0,3758
lazy.IBk	0,9118	0,0735	0,107
meta.AdditiveRegression	0,9295	0,0673	0,0937
meta.Bagging	0,9157	0,0768	0,103
meta.RegressionByDiscretization	0,9325	0,0615	0,0919

Tabela C.5: Métrica do mês de Agosto

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,6861	0,1462	0,1856
trees.RandomTree	0,7975	0,1111	0,1591
trees.RandomForest	0,8781	0,0965	0,1203
trees.M5P	0,9417	0,059	0,0846
functions.LinearRegression	0,7458	0,1378	0,1682
functions.MultilayerPerceptron	0,05	0,3557	0,4766
lazy.IBk	0,7184	0,141	0,1935
meta.AdditiveRegression	0,9088	0,0782	0,1038
meta.Bagging	0,6782	0,1475	0,1842
meta.RegressionByDiscretization	0,9113	0,07	0,1023

Tabela C.6: Métrica do mês de Setembro

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,7326	0,1252	0,1635
trees.RandomTree	0,7507	0,1303	0,1674
trees.RandomForest	0,864	0,0939	0,1188
trees.M5P	0,9346	0,0576	0,0841
functions.LinearRegression	0,8039	0,115	0,1412
functions.MultilayerPerceptron	0,1374	0,3531	0,4567
lazy.IBk	0,7168	0,1365	0,1759
meta.AdditiveRegression	0,9125	0,07	0,0961
meta.Bagging	0,7195	0,1284	0,1644
meta.RegressionByDiscretization	0,8791	0,0756	0,1127

Tabela C.7: Métrica do mês de Outubro

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,9224	0,0665	0,0959
trees.RandomTree	0,9053	0,71	0,1057
trees.RandomForest	0,9396	0,0609	0,0848
trees.M5P	0,9431	0,0511	0,0821
functions.LinearRegression	0,9234	0,0664	0,095
functions.MultilayerPerceptron	0,0175	0,4172	0,5219
lazy.IBk	0,8972	0,074	0,1109
meta.AdditiveRegression	0,9152	0,0665	0,0994
meta.Bagging	0,9076	0,0756	0,1042
meta.RegressionByDiscretization	0,9351	0,0599	0,0874

Tabela C.8: Métrica do mês de Novembro

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,8546	0,0937	0,1247
trees.RandomTree	0,8818	0,079	0,1137
trees.RandomForest	0,9272	0,0666	0,0899
trees.M5P	0,9411	0,0497	0,0801
functions.LinearRegression	0,9128	0,0685	0,098
functions.MultilayerPerceptron	0,1383	0,3033	0,4149
lazy.IBk	0,8279	0,0969	0,1382
meta.AdditiveRegression	0,9042	0,0683	0,1014
meta.Bagging	0,8464	0,0971	0,1262
meta.RegressionByDiscretization	0,9178	0,0652	0,0944

Tabela C.9: Métrica do mês de Dezembro

Algoritmos	**CC**	**MAE**	**RMSE**
trees.REPTree	0,9186	0,0756	0,105
trees.RandomTree	0,928	0,0688	0,0993
trees.RandomForest	0,9538	0,0587	0,0807
trees.M5P	0,9475	0,0512	0,0848
functions.LinearRegression	0,9538	0,0544	0,0797
functions.MultilayerPerceptron	0,0111	0,2976	0,3664
lazy.IBk	0,9233	0,071	0,1031
meta.AdditiveRegression	0,9321	0,0671	0,0963
meta.Bagging	0,9094	0,0826	0,1109
meta.RegressionByDiscretization	0,9244	0,0689	0,101

APÊNDICE D – CORRELAÇÕES MENSAIS DO FEIJÃO

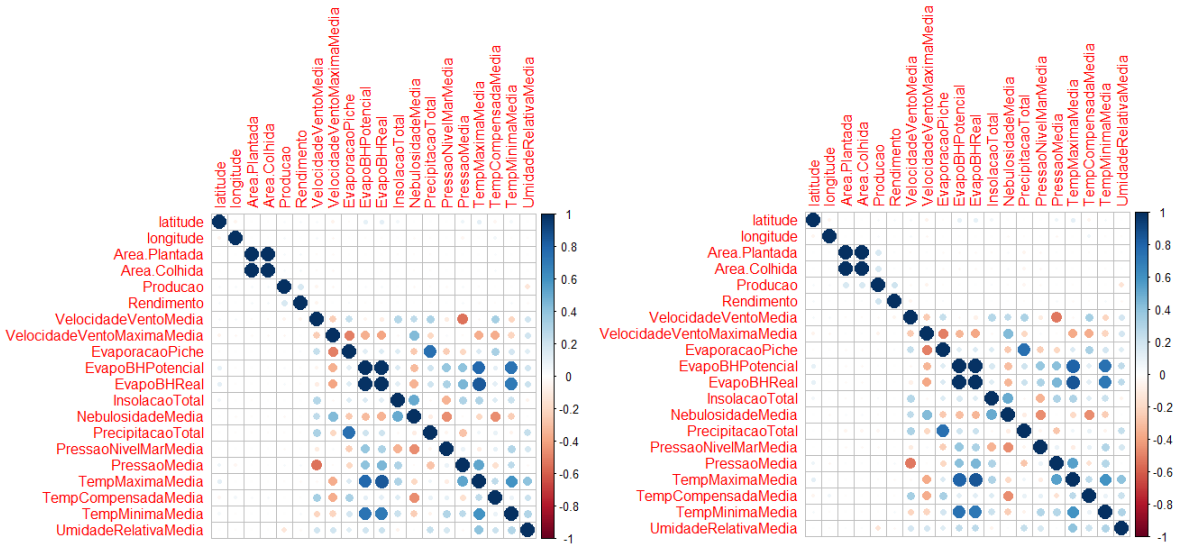


Figura D.1: Correlação do mês de Maio e Junho do feijão

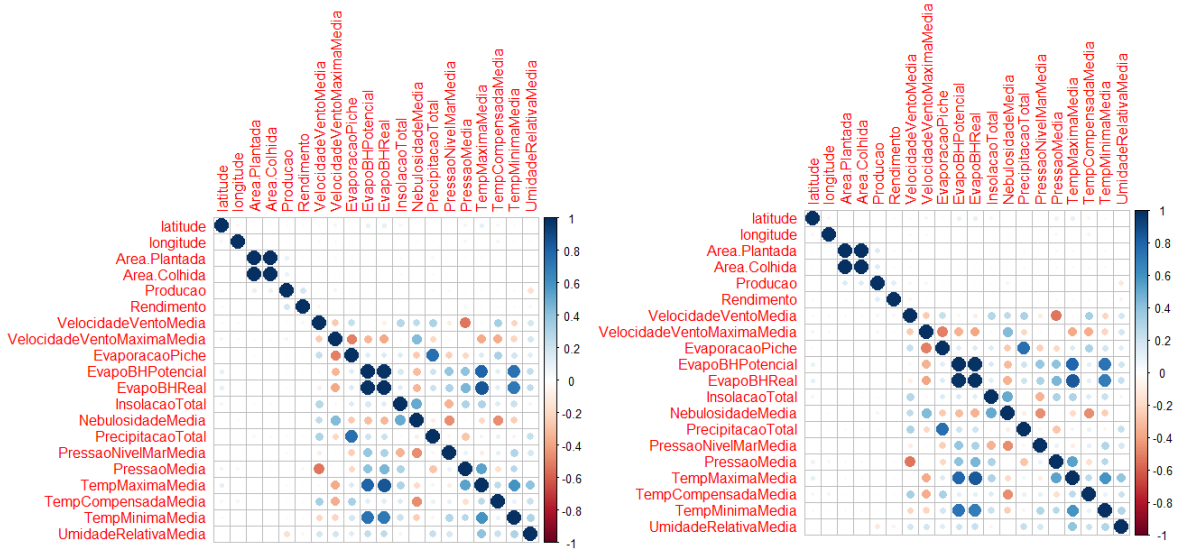


Figura D.2: Correlação do mês de Julho e Agosto do feijão

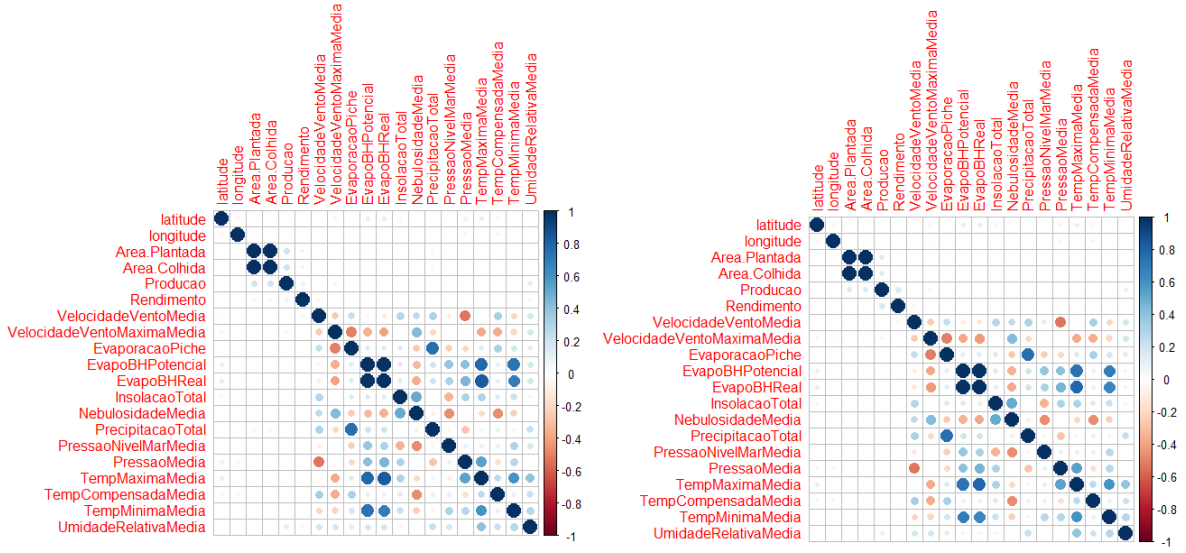


Figura D.3: Correlação do mês de Setembro e Outubro do feijão

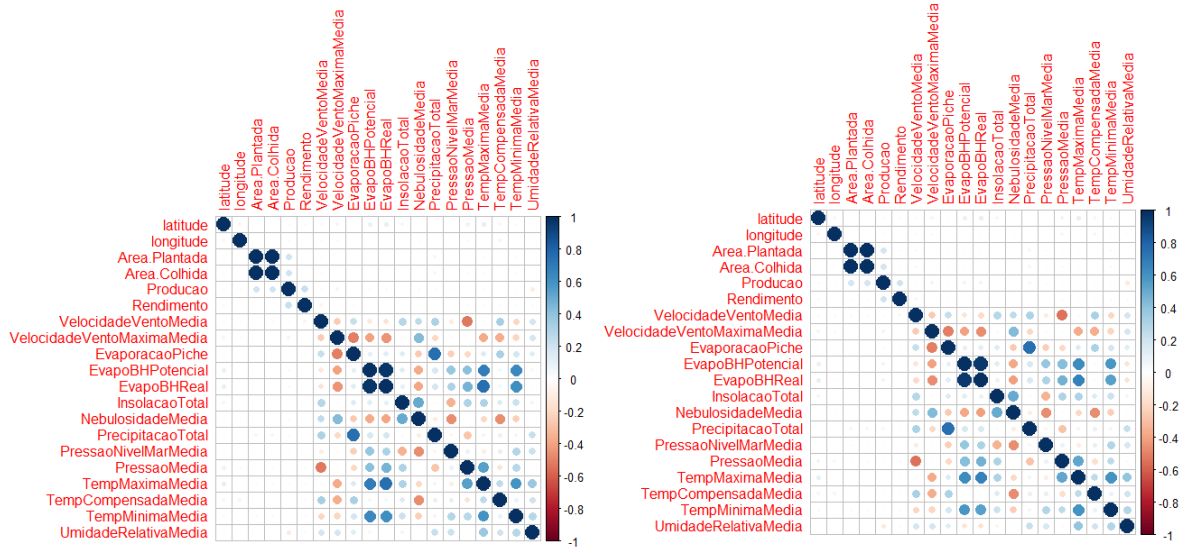


Figura D.4: Correlação do mês de Novembro e Dezembro do feijão

APÊNDICE E – TABELAS DAS MÉTRICAS DA CULTURA DO FEIJÃO

Tabela E.1: Métrica de Abril para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,7031	0,0793	0,109
trees.RandomTree	0,5798	0,0957	0,1315
trees.RandomForest	0,7368	0,0754	0,1013
trees.M5P	0,7044	0,0785	0,1092
functions.LinearRegression	0,7075	0,0797	0,1087
functions.MultilayerPerceptron	0,0212	0,1362	0,1681
lazy.IBk	0,6776	0,0873	0,1186
meta.AdditiveRegression	0,3483	0,1088	0,1403
meta.Bagging	0,689	0,0794	0,1065
meta.RegressionByDiscretization	0,6986	0,0799	0,1091

Tabela E.2: Métrica de Maio para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,3458	0,1099	0,1414
trees.RandomTree	0,427	0,1186	0,1489
trees.RandomForest	0,6532	0,0871	0,1116
trees.M5P	0,7482	0,0703	0,0971
functions.LinearRegression	0,5604	0,0988	0,1255
functions.MultilayerPerceptron	0,005	0,1383	0,1688
lazy.IBk	0,4003	0,1275	0,1587
meta.AdditiveRegression	0,5408	0,0914	0,1222
meta.Bagging	0,4678	0,1057	0,1341
meta.RegressionByDiscretization	0,1218	0,1119	0,1442

Tabela E.3: Métrica de Junho para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,6157	0,088	0,116
trees.RandomTree	0,5975	0,0907	0,1212
trees.RandomForest	0,7504	0,0709	0,0961
trees.M5P	0,1775	0,0814	0,4881
functions.LinearRegression	0,7119	0,075	0,1018
functions.MultilayerPerceptron	0,0136	0,1318	0,1615
lazy.IBk	0,62	0,0915	0,1224
meta.AdditiveRegression	0,4629	0,0958	0,1257
meta.Bagging	0,6048	0,0867	0,1153
meta.RegressionByDiscretization	0,6088	0,0885	0,1162

Tabela E.4: Métrica de Julho para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,683	0,0829	0,1148
trees.RandomTree	0,5843	0,0938	0,1337
trees.RandomForest	0,7295	0,077	0,1052
trees.M5P	0,714	0,0765	0,1098
functions.LinearRegression	0,6747	0,0839	0,1166
functions.MultilayerPerceptron	0,0254	0,148	0,185
lazy.IBk	0,6495	0,0918	0,1281
meta.AdditiveRegression	0,4313	0,1068	0,1377
meta.Bagging	0,669	0,0831	0,1151
meta.RegressionByDiscretization	0,6803	0,083	0,1146

Tabela E.5: Métrica de Agosto para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,4302	0,116	0,1471
trees.RandomTree	0,5593	0,1103	0,1453
trees.RandomForest	0,7784	0,0796	0,1047
trees.M5P	0,7827	0,0701	0,0994
functions.LinearRegression	0,7556	0,0758	0,1063
functions.MultilayerPerceptron	0,0762	0,1334	0,1678
lazy.IBk	0,4643	0,1276	0,1666
meta.AdditiveRegression	0,59	0,0991	0,1287
meta.Bagging	0,4962	0,1124	0,1425
meta.RegressionByDiscretization	0,4432	0,1096	0,1427

Tabela E.6: Métrica de Setembro para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,5524	0,0989	0,1298
trees.RandomTree	0,502	0,1072	0,1401
trees.RandomForest	0,7225	0,0801	0,1071
trees.M5P	0,7413	0,0716	0,1012
functions.LinearRegression	0,639	0,0873	0,1188
functions.MultilayerPerceptron	0,0252	0,1659	0,2264
lazy.IBk	0,4255	0,1232	0,1626
meta.AdditiveRegression	0,5115	0,0996	0,1286
meta.Bagging	0,5362	0,0988	0,1299
meta.RegressionByDiscretization	0,3366	0,1115	0,1434

Tabela E.7: Métrica de Outubro para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,6219	0,0854	0,1136
trees.RandomTree	0,4943	0,0991	0,1313
trees.RandomForest	0,6768	0,0764	0,1034
trees.M5P	0,694	0,0752	0,1029
functions.LinearRegression	0,6673	0,08	0,1072
functions.MultilayerPerceptron	0,0012	0,1707	0,2268
lazy.IBk	0,5498	0,1002	0,1331
meta.AdditiveRegression	0,418	0,0969	0,1269
meta.Bagging	0,6096	0,0848	0,1133
meta.RegressionByDiscretization	0,6157	0,0851	0,1136

Tabela E.8: Métrica de Novembro para a cultura do feijão

Algoritmos	CC	MAE	RMSE
trees.REPTree	0,7185	0,0758	0,1053
trees.RandomTree	0,6031	0,0923	0,1255
trees.RandomForest	0,7544	0,0718	0,0979
trees.M5P	0,7227	0,0748	0,1046
functions.LinearRegression	0,7257	0,0749	0,1042
functions.MultilayerPerceptron	0,0169	0,1334	0,1656
lazy.IBk	0,6629	0,0865	0,1212
meta.AdditiveRegression	0,4068	0,1037	0,1351
meta.Bagging	0,7045	0,0763	0,106
meta.RegressionByDiscretization	0,7132	0,0761	0,1056

Tabela E.9: Métrica de Dezembro para a cultura do feijão

Algoritmos	**CC**	**MAE**	**RMSE**
trees.REPTree	0,631	0,085	0,1154
trees.RandomTree	0,5895	0,0916	0,1251
trees.RandomForest	0,7386	0,072	0,0987
trees.M5P	0,722	0,0701	0,1015
functions.LinearRegression	0,6916	0,0764	0,1068
functions.MultilayerPerceptron	0,0238	0,1324	0,1608
lazy.IBk	0,6269	0,0908	0,1237
meta.AdditiveRegression	0,4544	0,0982	0,1276
meta.Bagging	0,6212	0,0842	0,1146
meta.RegressionByDiscretization	0,6268	0,0851	0,1153

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br